UNIVERSITÉ DE STRASBOURG

DOCTORAL SCHOOL (MSII)

ICube Laboratory (UMR 7357)

Research Group CAMMA on

Computational Analysis and Modeling of Medical Activities

Thesis presented by

Vinkle Kumar Srivastav

on

19th November 2021

for obtaining the degree of **Doctor of Philosophy**

From the **University of Strasbourg** delivered by the **doctoral school MSII**

Unsupervised Domain Adaptation Approaches for Person Localization in the Operating Rooms

Thesis Directors

Prof. Nicolas Padoy
Professor of Computer Science,
University of Strasbourg,
IHU Strasbourg, France

Prof. Afshin Gangi
Professor of Radiology,
CHU Strasbourg, France

Chair of the Committee

Prof. Michel de Mathelin
Professor of Robotics,
University of Strasbourg, France

Thesis Examiners

Prof. Gregory Hager
Professor of Computer Science,
Johns Hopkins University, USA

Prof. Slobodan Ilic
Professor of Computer Science,
Technical University of Munich (TUM), Germany

Dr. Vasileios Belagiannis
Professor of Computer Science,
University of Ulm, Germany

Abstract

The recent emergence of surgical data science holds the promise to enable a new generation of operating room (OR) support systems. The fine-grained localization of clinicians in the OR, either at the keypoint-level using human pose estimation or at the pixel-level using instance segmentation, is a key component to design such systems. The task is however challenging not only because OR images contain significant visual domain differences compared to traditional vision datasets, but also because data and annotations are hard to collect and generate in the Operating Room (OR), due to privacy concerns. Approaches that can adapt a model to an unseen and unlabeled target domain are therefore very promising.

In this dissertation, we explore Unsupervised Domain Adaptation (UDA) methods to enable visual learning for the target domain, the operating room, by working in two complementary directions. First, we study how low-resolution images with a downsampling factor as low as 12x can be used for fine-grained clinicians localization to address privacy concerns. Second, we propose several self-supervised methods to transfer learned information from a labeled source domain to an unlabeled target domain to address the visual domain shift and the lack of annotations. These methods employ self-supervised predictions in allowing the model to learn and adapt to the unlabeled target domain. We first propose to perform domain adaptation across visual modalities of color (RGB) to depth (D) images by exploiting synchronized properties of the RGB-D images and utilizing state-of-the-art human pose estimation models on RGB images for human pose estimation on depth images. Second, we explore *knowledge-* and *data-distillation* to generate accurate pseudo labels from a multi-stage, larger, and accurate teacher model to train a single-stage, smaller, and deployable student model for joint 2D/3D human pose estimation. Finally, we propose to employ spatial and geometric constraints on the different augmentations of the image and disentangle feature normalization layers in the backbone model to simultaneously learn from the labeled source and the unlabeled target domain data for joint pose estimation and instance segmentation. To demonstrate the effectiveness of our proposed approaches, we release the first public dataset, called the multi-view operating room (*MVOR*), generated from recordings of real clinical interventions. We obtain state-of-the-art results on the *MVOR* dataset,

specifically on the privacy-preserving low-resolution OR images. We hope our proposed UDA approaches could help to scale up and deploy novel AI assistance applications for the OR environments.

Acknowledgments

Although, on paper, I get to take the credit as the sole author of this dissertation; however, in reality, this is a purely collaborative endeavor. This dissertation would not have been possible without the support and guidance of my supervisors, mentors, colleagues, friends, and family. I would like to start by expressing my sincere gratitude to my Ph.D. advisor, Nicolas Padoy, for taking me in as their advisee.

I want to thank Nicolas for the opportunity to work in a uniquely collaborative environment where clinicians and computer scientists can work together to solve clinically relevant problems. Nicolas has taught me how to think deeply about a given research problem yet not lose focus on the long-term goal. Offering selfless care, writing simple, concise yet compelling sentences, and arguing with reason are some of the invaluable qualities I learned from him. Despite leading a multidisciplinary and constantly evolving large group, Nicolas has a unique ability to take excellent care of his Ph.D. students. I am thankful for his continued support and guidance in making me the researcher I am today. This dissertation would not be possible without the support of Prof. Afshin Gangi, my co-supervisor, who has always been supportive and kind, despite his busy schedule as the head of the interventional imaging department. I am extremely thankful to him for allowing us the record the OR data from the real clinical interventions and for helping us create and publish the MVOR dataset.

I am also extremely grateful to have Gregory Hager, Slobodan Ilic, Michel de Mathelin, and Vasileios Belagiannis as my thesis committee members. It is an honor to get feedback on my dissertation from the pioneers of surgical data science. Their contributions to surgical data science have made the field what it is today. The research ideas proposed in this dissertation are motivated by their research principles. I am thankful to them for taking the time to read my dissertation and critically evaluate it.

I would like to thank our former Ph.D. students of the CAMMA group, Rahim Kadkhodamohammadi, Andru Twinanda, and Nicolas Loy. Their extreme hard work in collecting and generating the OR datasets has provided us with invaluable resources to expand our research. I am specifically grateful to Rahim for guiding me during the initial year of my Ph.D. His inspiring work in the human pose estimation from the external OR videos has given me unique perspectives in starting my research.

Working alongside my Ph.D. colleagues, Chinedu Nwoye and Tong Yu, was great fun and learning experience. Chinedu, with his fruitful punchy remarks, and Tong, with his selfless support, have enormously helped me navigate my Ph.D. journey. I want to thank CAMMA group members and interns for various fun and research activities. Working and collaborating with Thibaut was a fruitful learning experience. My memories at CAMMA will not be complete without Pietro Mascagni, Deepak Alapatt, Sanat Ramesh, Luca Sestini, Armine Vardazaryan, Gaurav Yengera, Alexandre Krebs, Cindy Rolland, Antoine Fleurentin, Jean-paul Mazellier, Georgios Exarchakis, Alexandros Karargyris, Idris hamoud, and Isabella Bolognese. Hanging out with them has often helped me not get homesick. Thanks to IHU Strasbourg for providing me with a stimulating and welcoming environment.

I would like to thank my brothers, Amit Shrivastav, Sumit Shrivastav, and my sister Shilpy, for their continued help and support. Their support, especially during challenging COVID times, helped me maintain my sanity. I also like to thank my friends, Jarnail Singh, Pankaj Yadav, Ravi Pal, Ripul Jain, Britty Baby, Ramandeep Singh, Chandrika, Himanshu Gandhi, Aruna, and Vijay, for their help and support.

Finally, I would like to thank my parents, Sh. Dinesh Prasad and Smt. Savitri Devi. If not for their unconditional love and unlimited support, I would not have been fortunate enough to be the first engineer and first doctorate in our family tree. Their extreme hard work and uncountable sacrifices have provided me the opportunities to reach where I am today.

I gratefully acknowledge the financial support which funded my work, received from French state funds managed by the ANR within the Investissements d'Avenir program under references ANR-16-CE33-0009 (DeepSurg) and ANR-10-IAHU-02 (IHU Strasbourg). I would like to acknowledge the university of strasbourg support for the HPC cluster and HPC resources of IDRIS under the allocation 20XX-[AD011011631R1] made by GENCI.

To my parents, Sh. Dinesh Prasad and Smt. Savitri Devi

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AE Associative Embedding. 24

AI Artificial Intelligence. 4

AP Average Precision. 42, 72

ASR Automatic Speech Recognition. 19, 20

CAI Computer-Assisted Intervention. 9

CNN Convolutional Neural Network. 107, 141

CP Cerebral Palsy. 100

DConv Deformable Convolution. xix, 41

DFN Disentangled Feature Normalization. 14

EMA Exponential Moving Average. 21

FPN Feature Pyramid Network. xix, 41

GAN Generative Adversarial Network. 22

GPU Graphics Processing Unit. 4

HDD Histogram of Depth Difference. 25

HPE Human Pose Estimation. 12, 13, 18, 23, 24, 26, 39, 48, 49, 60

ICU Intensive Care Unit. 48

IoU Intersection over Union. 39, 42

meanPCK Mean Percentage of Correct Keypoints. xix, 45

MIS Minimally Invasive Surgical. 9

MPJPE Mean Per Joint Position Error. 42

MVOR Multi-view Operating Room. 32, 104, 105

NLP Natural Language Processing. 19, 20

OKS Object Keypoint Similarity. 42

OR Operating Room. i, ii, viii, xv, 5, 8, 11, 26, 32, 48, 60, 98, 99, 103, 104, 106, 108, 114, 118, 142

ORPM Occlusion-robust Pose-maps. 25

RGBD Red-Green-Blue-Depth. 13, 104, 107, 134, 141

RPN Region Proposal Network. 38

SSL Semi-supervised Learning. 21, 72, 105

UDA Unsupervised Domain Adaptation. i, ii, 7, 11, 13–15, 19–21, 49, 70, 99, 100, 104, 105, 107, 108, 132, 141, 142

UMAP Uniform Manifold Approximation and Projection for Dimension Reduction. xi, xvi, 12, 130

Part I
Introduction and related work

1 Introduction

What we want is a machine that can learn from experience.
– Alan Turing, 1948

The large amount of data produced by the modern operating rooms can enable the development of a new generation of support systems, such as "surgical control towers" and "OR black-boxes" for real-time activity analysis and efficient offline recording, respectively, with the overall aim to improve patient care [Mascagni 2021c].

Chapter Summary

1.1 Background

One of the theories of evolution suggests that we are living through the fourth epoch of technological advancement: the information age [Kurzweil 2005], witnessing an unprecedented rise in computing power and digital data. Compute capabilities of digital devices have seen an incredible increase over the past few years. Much of it can be attributed to the current Graphics Processing Unit (GPU) whose compute power can be equivalently compared with the computers that navigated the first satellite launch into orbit and the moon landings. If we follow Moravec's argument [Moravec 1998], a single computer will have the same computing power as humans by 2025. Alongside computing power, the past decade has also witnessed an exponential growth in digital data facilitated by the rise of the internet and inexpensive visual sensors. According to a rough estimates, we upload 400 million pictures every day on Facebook and 300 hours of video content every minute on YouTube. To put this number in perspective, Facebook adds a full ImageNet sized dataset [Russakovsky 2015] - one of the largest computer vision datasets with 14 million images divided among 22k categories - every hour, and YouTube does this every 13 minutes.

This rapid growth in computing power and large-scale datasets has brought about a renaissance in Artificial Intelligence (AI), specifically in 'deep learning': a subset of machine learning techniques capable of learning generic representations from the raw data. The deep learning algorithms aim to estimate the parameters of deep neural networks from the pair of input and desired output examples; it does so by iterating through billions of input-output possibilities in an optimization loop. Effectively training these algorithms requires massive datasets needed to capture virtually all the different possibilities. Although the foundational concepts of deep learning have been around

since the 1980s [Rumelhart 1986, LeCun 1989], it is the current highly efficient computers and the large-scale datasets that are unlocking the true AI capabilities.

One component in the large-scale datasets contributing to the success of AI is the need for human-generated labels. These labels provide necessary supervision signals to let the deep neural network converge towards the optimal solution. It was first illustrated through the state-of-the-art performance for the image-classification task on the ImageNet dataset [Russakovsky 2015, Krizhevsky 2012]. This flagship benchmark result exploiting the large-scale manually labeled dataset started the successful journey of deep neural networks. The supervised deep learning paradigm has now flourished to more complex computer vision tasks with improved performance and new applications. For example, Mask-RCNN [He 2017] can provide pixel-level segmentation for the objects in the image; OpenPose [Cao 2017] can localize the body joints of all the persons in a given image.

However, these models work well if the test time images have the same distribution as the train time image. More often, the images a model receives at test time differ significantly from the ones received at training time leading to the failure of such models in a real-world deployment setting. To illustrate this, we evaluate one of the state-of-the-art models, ResNet-50 [He 2016], on some YouTube images. Although these images belong to the same ImageNet classes, the model fails remarkably on these real-world data, see figure 1.1. Similarly, when we evaluate state-of-the-art human pose estimation models on the real-world OR data, we see significant localization errors, as shown in the figure 1.2. The problem is nevertheless not new and even goes back to the inception of AI. An often-told story from the early 1970s goes as follows: ARPA (now known as DARPA (Defense Advanced Research Projects Agency)) organized a challenge to classify a given image in two classes: images containing tank vs. non-tank. They gave participants a dataset for this binary image classification problem. Although the AI model achieved accurate results on the given dataset, the model failed remarkably when deployed in the real-world setting. Further analysis showed that the dataset contained images of the tank on sunny days and images of non-tank on cloudy days. So instead of learning about the visual concept of a tank, the AI models found a shortcut and classified the images based on just the brightness [Efros 2019].

One way to overcome the challenge of train-test distribution mismatch is to train the model with manual labels from all types of domain distributions. The hope is that as the model would have seen different possibilities, the test-time generalization would eventually become an interpolation problem. However, it poses challenges in the following two dimensions: first, manual annotation is time-consuming and expensive; for example, the ImageNet dataset took about 19 years for the annotations, and for more involved annotation task such as pixel-level segmentation, annotating a single image can take up to 90 minutes [Cordts 2016]. Second, the manual annotation gets further complicated if the target domain is privacy-sensitive or requires expert annotations.

One such domain is healthcare which actively faces the challenges of lack of manually annotated data and the need for privacy-preserving approaches. In this sector, surgery

Figure 1.1: Qualitative evaluation of the ResNet-50 model on real-world frames. The model fails and misclassifies, for example, "Oxygen mask" to detect classes like "syringe," "Stethoscope," and "Barbershop." See [Williams 2011] for more examples.

Figure 1.2: Qualitative evaluation of the state-of-the-art 2D human pose models show large localization errors on the real-world frames from the OR. We evaluate OpenPose [Cao 2017], Keypoint-RCNN [He 2017], CPN [Chen 2018a], and AlphaPose [Fang 2017] models.

in an operating room Operating Room (OR) is an exciting and challenging field, given its socio-economic value. Clinicians in the modern-day OR interact in a constrained and cluttered environment to achieve one common goal: to heal a patient. Accurate localization of the clinicians from the ceiling-mounted cameras in the dynamic OR environment could open up a path to various new applications ranging from augmented

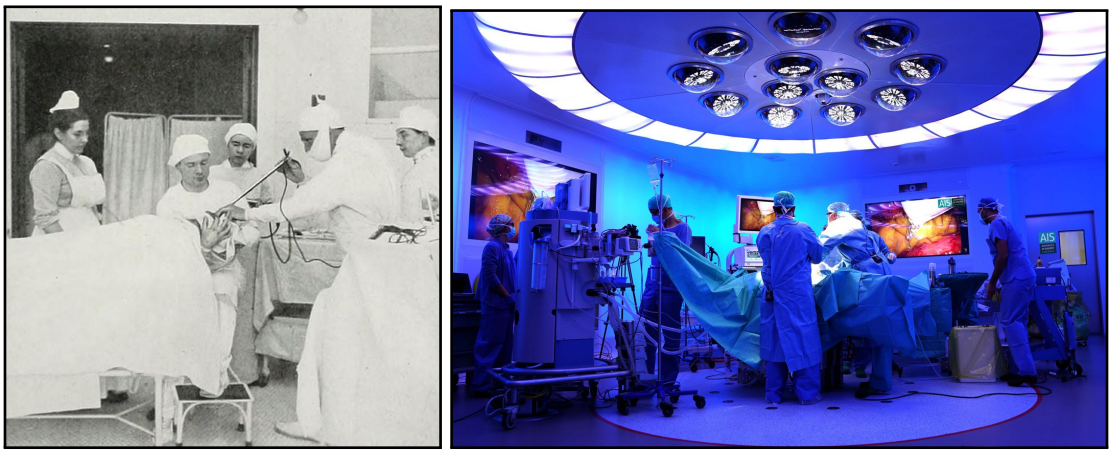


Figure 1.3: The figure shows an illustrative comparison of an OR from the early 1900 with a modern-day OR. The left-hand side, image courtesy [Jackson 1915], shows a surgeon performing an esophagoscopy procedure from limited knowledge of anatomy and using basic instruments. The modern-day OR shown in the right-hand side, image courtesy [5Gw], has become very specialized where a dedicated team of surgeons and the clinical staff uses advanced technology to perform surgery while minimizing damage to the intricate patient anatomy.

reality, automatic skills evaluation to novel context-aware systems.

This thesis studies the problem of Unsupervised Domain Adaptation (UDA) in the context of the OR. We propose several domain-adaptation approaches for person localization without using any manual annotations from the OR. We also consider the privacy-sensitive nature of the OR to extend further the methods to respect the privacy of clinicians and patients. In the following, we discuss the OR as a target domain and describe various applications of the person localization problem, followed by a brief description of our UDA approaches.

1.2 Operating room

The OR is a dedicated unit inside the hospital where clinicians collaborate in a complex, high-risk, dynamic, technologically advanced, and cluttered environment to treat the patients by altering their anatomy. The history of the surgery is much older, even date back to 3000 BC when the Egyptians performed the first recorded *trephination* - a surgical intervention where a hole is drilled into the skull using simple surgical tools [Rutkow 2000]. Much of the advances in surgery have been brought in the last century predominantly due to the advances in medical technology ranging from cutting-edge surgical tools, navigation and monitoring systems to novel imaging technologies. The figure 1.3 illustrate how a modern-day OR has evolved compared to an OR from the last century. Coupled with the digital revolution, the OR has now become a financial nexus inside the hospital. A rough estimate per-minute cost of the OR comes around $36 and accounts for up to 40% of a hospital's costs and 60-70% of revenues [Childers 2018].

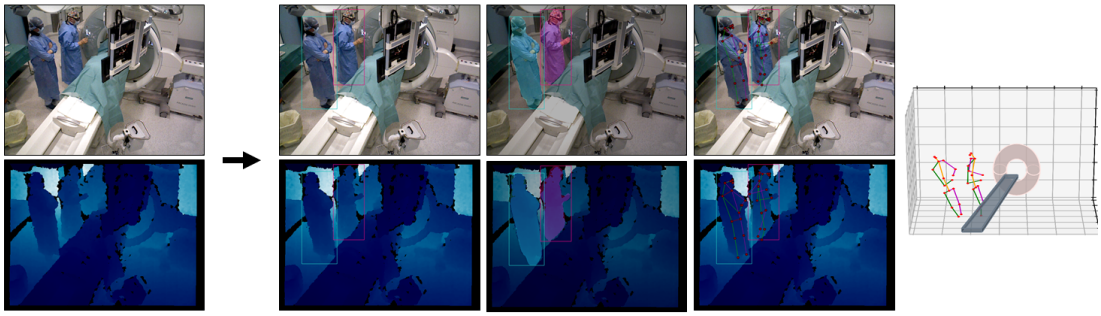


Figure 1.4: The person localization can be performed at the coarse level using bounding box detection or at the fine-grained level using pixel-based instance segmentation or keypoint-based 2D/3D human pose estimation.

1.2.1 Privacy in the OR

The current age of AI exploiting the "big data" has started to enable intelligent applications with improved efficacy in various sectors while simultaneously raising the growing public concern regarding the ethical use of data. Indeed, the recent controversies [Powles 2017] have raised public awareness regarding how personal data should be collected and controlled, along with how the AI algorithms should use personal data in a privacy-safe way[1] [Symons 2017]. Health-care data, especially the direct video recording of OR using ceiling cameras, raises the understandable concern about misuse of this highly privacy-sensitive OR data. Adapting a model to very low-resolution images has been suggested in the literature to improve privacy [Chou 2018]. Indeed, as low-resolution images significantly degrade the spatial details, it could provide a viable means to improve privacy. However, fine-grained spatial localization tasks such as pose estimation or instance segmentation become challenging. This dissertation has explored the ways to address these concerns by utilizing spatial and geometric constraints to effectively adapt the model to privacy-preserving low-resolution OR images.

1.3 Person localization in the OR

The OR produces rich signals using various instruments and sensors to monitor and treat the patient and document the procedure. Notably, ceiling cameras capture a global view of the OR in the form of color, depth, or both types of images. We can use these non-invasive informative signals for a variety of new applications. As the clinicians in the OR are the main dynamic actors, these signals could help for their localization. The clinicians' localization can be performed at the coarse level using bounding box detection or at the fine-grained level using human pose estimation or person instance segmentation, see figure 1.4. In the following, we discuss applications of various localization approaches in the OR.

[1] https://decodeproject.eu/

1.3.1 Applications of person localization in the OR

1.3.1.1 Context-aware system

A context-aware system is a system that can understand different activities and provide an appropriate response in a given environment. In the OR, surgery is the main activity performed by clinicians in a visually cluttered environment consisting of complicated steps and proceeds in progressive stages. A context-aware system for the OR therefore aims to automatically track and analyze progress in an ongoing surgical operation.

Digitally enabled modern OR provides multi-modality signals consisting of videos captured from either endoscopic or ceiling-mounted cameras, radiographic images, device signals, and electronic health reports.

The tremendous progress in AI has enabled the Computer-Assisted Intervention (CAI) community to exploit these rich signals to develop various components for the context-aware systems for the OR. Inspired from other high-stake sectors such as aviation and Formula One [Helmreich 2000, Gawande 2011, Catchpole 2007], on the one hand, these components can serve as OR black-box [Goldenberg 2017], similar to flight recorders, to capture the multi-modality OR data for offline analysis. On the other hand, these components can serve as a surgical control tower [Padoy 2018], similar to air traffic control towers, to online stream the OR data for overseeing, coordinating, and providing an instantaneous assessment of the OR activities [Mascagni 2021c]. A variety of underlying components have been developed by analysing endoscopic videos. Example of such components are surgery type recognition [Kannan 2019], surgical phase and activity recognition [Yengera 2018, Yu 2018, Ramesh 2021, Padoy 2009, Padoy 2012, Twinanda 2015, Tran 2016, Twinanda 2016a], surgical tool detection and tracking [Bouget 2015, Nwoye 2019, Sestini 2021], assessment of critical view of safety in laparoscopic cholecystectomy procedures [Mascagni 2021b, Mascagni 2021d, Mascagni 2021a, Mascagni 2021e], and fine-grained action triplet recognition [Nwoye 2020].

While endoscopic videos provide a rich context for the surgical workflow analysis in minimally invasive surgical procedures, they do not capture the activities happening inside the whole OR. Therefore, using ceiling-mounted cameras to capture the scene from an external view can help recognize and analyze activities by localizing clinicians in the OR. As the clinicians are the main dynamic actors in an otherwise passive OR environment, localizing and monitoring their postures would endow the machines using a variety of signals to build a complete context-aware system [Nara 2010, Agarwal 2007, Meißner 2014, Bardram 2011].

1.3.1.2 Surgical skill assessment

Minimally Invasive Surgical (MIS) procedures have gained enormous interest in the past decade due to reduction in postoperative recovery time, morbidity, hospitalization time, and cost of patient care [Jaffray 2005]. It provides a surgeon with high-fidelity visualization of the complex surgical site while minimizing damage to the intricate

anatomy. A surgeon in minimally invasive surgical procedures needs to perform surgery by looking at the two-dimensional screen; the margin of error in these procedures is minimal. It requires elaborate and effective training for eye-hand coordination, depth perception, and bi-manual dexterity. As the complexity in these technologically advanced surgical procedures has been increasing, iatrogenic errors have also been drawing increased attention to the skills of a surgeon [Donaldson 2000, Makary 2016]. The notion of "learning by doing" is diminishing due to high-risk factors in surgery, increased patient demands, and scrutiny on a surgeon's performance [Bridges 1999, Vozenilek 2004]. Therefore, designing effective methods for teaching and assessing surgical skills is imperative in the hospital. A study in [Wanzel 2002] suggests that direct supervision of an expert surgeon on the gestures and movements of novice surgical residents can improve their performance. An expert surgeon directly supervising the novice surgeons for all the repeatable training procedures would be a time-consuming, costly and non-scalable process [Ghani 2016]. Authors in [Reiley 2011] suggest that the motion of the tool and the body joints of the trainee surgeons can provide objective and measurable parameters to evaluate surgical skills. Given the repeated nature of the training procedures and the advent of vision-based AI approaches, an automatic skills evaluation could provide an objective assessment, real-time feedback, and staged development of skills without the supervision of an expert surgeon [Bridges 1999]. Human pose estimation utilizing the non-obstructive camera feeds to track and analyze the body joints could be a fundamental step towards building such automatic evaluation systems.

1.3.1.3 Radiation safety monitoring

Intraoperative X-ray imaging has become fundamental in several fields of medicine, especially in hybrid surgeries such as minimally invasive image-guided procedures. These intra-operative procedures rely on the x-rays to control and monitor the tools inserted in the patient; hence the clinicians need to remain close to the patient and eventually to the potentially harmful ionizing radiations. The clinicians usually wear protective lead shields and use dosimeters to monitor the exposure. As these harmful radiations are invisible to the eye, and studies have consistently shown their deteriorating effects, especially on the clinicians [Vanhavere 2008, Carinou 2011, Roguin 2013, Nowak 2020], it is crucial to develop an intuitive radiation risk awareness system for the hybrid OR. Recent works such as [Ladikos 2010, Rodas 2016, Krebs 2021] have developed a radiation simulation system that uses person-body models to measure the radiations at different body parts. These systems, which show great promise to build such an intuitive radiation risk awareness in the OR, can be further extended by accurately measuring pixel-level radiation risks using fine-grained localization such as person instance segmentation.

1.3.1.4 Modeling team dynamics

Humans convey their thoughts, emotions, and intentions using a variety of signals, for example, language, voice, facial expression, and body gestures. Clinicians in the OR also use the same signals to communicate in the OR, either verbally or non-verbally. Non-verbal communication in the OR is essential in a critical phase of surgery to work effectively. Endowing machines with the ability to encode and decode these broad spectrum of human gestures would facilitate understanding interactions, non-verbal communications and cognitive load, especially in the critical phases of surgery [Dias 2019, Soenens 2021].

1.3.2 Challenges for person localization in the OR

As discussed in section 1 and illustrated in the figure 1.1, a change in the data distribution could result in a significant failure of the recognition models. OR as a target domain poses specific challenges in data distribution changes at the global and person instance levels. The OR has particular lighting conditions giving global level appearance changes. The clinicians wear loose clothes and surgical masks and occlude one another due to close proximity and instrument clutter, giving instance-level appearance changes. As the loose and texture-less clothes worn by clinicians appear very similar to the materials used to cover the other surfaces in the room, it becomes particularly challenging for the models trained on the natural images to generalize in the OR environment. Figure 1.5 shows global and instance-level visual differences between natural images and OR images, and how a model trained on the natural images fails in the challenging OR scenario. One way to overcome such domain differences is to fine-tune a model on the manually labeled data from the target domain. This however is particularly unscalable for the OR due to privacy concerns. The scalable and successful crowd-sourcing platforms, for example, Amazon Turk, can not be easily used for the privacy-sensitive OR images to provide large-scale manually labeled data. Approaches that can adapt a model to the unseen and unlabeled target domain are therefore very promising.

1.4 Our approach

This thesis explores several UDA directions to enable the visual learning on the OR images as a target domain while simultaneously tackling OR privacy. The aim is to adapt learned information from a labeled source domain to the unlabeled target domain sharing a common label space. We further propose initial works that take into consideration the privacy-sensitive nature of the OR environment. The repercussions of our approaches are twofold; first, as manual annotations, especially for pixel-based dense localization tasks, are considered the main bottleneck in the design of AI systems, a model trained on just the unlabeled data are easily scalable to more target domains. Second, as the approaches consider the privacy-sensitive nature of the target domain, it could be better accepted in the clinical institutions.

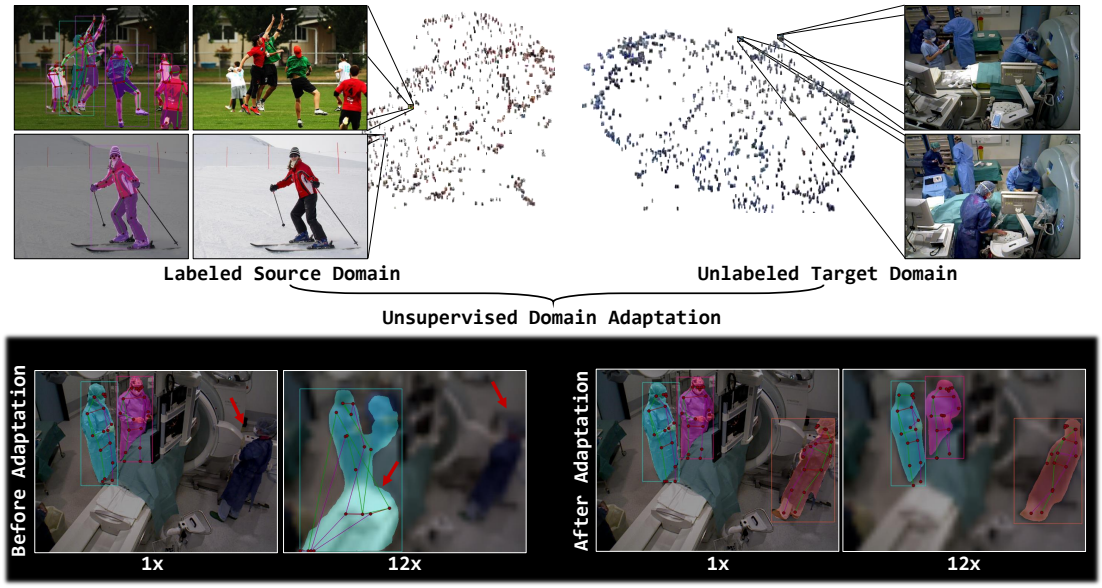


Figure 1.5: Global and instance-level visual differences between *source domain* natural images and *target domain* OR images. When a model trained on the source domain is applied to the unseen target domain, we see a substantial decrease in the localization accuracy and increase in the missed detections. Unsupervised domain adaptation approaches aim to effectively adapt a given source domain trained model to the target domain. The separate clusters of the source domain and the target domain images are obtained by running a dimension reduction technique: Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) [McInnes 2018, Duhaime]. The source and the target domain images are a subset of COCO [Lin 2014] and the MVOR [Srivastav 2018] datasets, respectively.

How can a model trained on fundamentally different domain data adapt itself to another target domain by just using unlabeled target domain data for supervision? We hypothesize that the use of self-supervised predictions is the key to this answer. As suggested in the literature ranging from centuries-long human philosophy to modern neuro-science, the human brain follows a similar idea to refine its predictions in an unsupervised way. David Hume, a famous philosopher from the 1800s, emphasizes in his book *Treatise of Human Nature* that mental perceptions - *ideas* or *representations* - from the sensory data - *impressions* - enforces invariant associations for the semantically similar images [Norton 2000]. Modern neuro-science has also observed experimentally that self-supervised deep-learning models to learn generic features representations resembles the ventral visual stream of the brain [Konkle 2021]. Moreover, the recent results in the self-supervised learning methods to learn generic feature representations from unlabeled data uses the similar idea that different views of the same image under different data augmentation should give similar predictions [Chen 2020, Grill 2020, He 2020]. Inspired by these ideas, we propose generic mechanisms that employ prediction as a self-supervisory signal in allowing the model to learn and adapt to the target domain. In the following, we briefly describe our contributions.

I. Release of *MVOR* dataset and comparison of state-of-the-art methods

As our first contribution, we introduce a new multi-view operating room dataset, called *MVOR*, as the first public dataset recorded during real clinical interventions for multi-person detection and 2D/3D Human Pose Estimation (HPE). *MVOR* shows the inherent visual challenges from the real-world OR environment, illustrating significant variations in color distributions compared to natural images and clinicians wearing loose clothes and masks under close proximity. We release the *MVOR* dataset with the aim to advance the state-of-the-art for fine-grained person localization in the OR by proposing it as a comparative *test-bed*. We evaluate state-of-the-art approaches for person detection and human pose estimation at the original (1x) and downsampled (12x) scales. We observe a significant performance degradation, specifically on the low-resolution images helping us to set up initial *source-only* baselines for our proposed UDA approaches.

II. Domain adaptation across visual modalities for HPE on low-resolution depth images

As our second contribution, we design UDA approaches for HPE on low-resolution depth images. As the depth images are texture-less and only encode the distance between an object to the sensor, these provide a viable option to preserve the privacy of patients and clinicians. As highlighted in 1.2.1, we further use only the low-resolution depth images at the test time to enforce more substantial constraints for the privacy-sensitive OR environment.

To train a model on the low-resolution depth images without manual annotations, we put forward an idea that two different visual modalities, such as color and depth images, can serve as two distinct domains. If these two domains are synchronized, as simply possible through the Red-Green-Blue-Depth (RGBD) cameras, then a model working reasonably well on one domain can effectively be adapted to the other. To enable this, we propose to perform inference on the color images using state-of-the-art HPE models, refine the inference results to generate pseudo labels, and transfer the pseudo labels to the corresponding depth image for the training.

We further propose two training strategies using the generated pseudo labels to adapt a model to the low-resolution depth images. As the first strategy, we propose to integrate super-resolution feature maps in the bottom-up RTPose [Cao 2017] method that utilizes intermediate super-resolution feature maps for effective learning of the high-frequency features. As the second strategy, we exploit advanced data-augmentations such as low-resolution down- and up-sampling, rand-augment [Cubuk 2020] and random cut-out [DeVries 2017] in the top-down Keypoint-RCNN [He 2017] model. We show significantly better results on the challenging *MVOR* dataset for both of our strategies, specifically on the privacy-preserving low-resolution depth images with a downsampling factor as low as 12x.

III. Self-supervision on unlabelled OR color images for joint 2D/3D human pose estimation

For our third contribution, we propose a UDA approach based on the *teacher-student* learning paradigm to develop an easily deployable model for joint 2D/3D human pose estimation on the OR color images. The teacher model exploits knowledge-distillation [Hinton 2015, Zhang 2019a] - using complex three-stage models - along with data-distillation [Radosavovic 2018] to generate accurate pseudo labels. We propose to use two sets of labels from the teacher model: a *hard-labels* set by removing low confidence detections and a *soft-labels* set by keeping all the detections along with their confidence value.

We further propose an end-to-end single-stage student model based on Mask-RCNN [He 2017] where we replace mask-head with a keypoint-head for joint 2D and 3D pose estimation. The student model exploits both the *hard-labels* and the *soft-labels* for effective training. Furthermore, to adapt the model to the privacy-preserving low-resolution images, we extend the data augmentation pipeline to generate low-resolution images by down-sampling and up-sampling the input OR image with a random scaling factor between 1x to 12x. The model trained on these very low-resolution OR images learns to give accurate results as the training progresses. Results on the *MVOR* dataset show that the student model performs on par with the teacher network despite being a lightweight and single-stage. Furthermore, it can also yield accurate results on low-resolution images, as needed to ensure privacy, even using a downsampling rate of 12x.

IV. Unsupervised domain adaptation for clinician pose estimation and instance segmentation in the OR

In our first two contributions, we propose to use a robust multi-stage teacher model to generate accurate pseudo labels. However, a strong teacher may not always be available to train a student model. In this work, we ask the following question: instead of relying on a strong teacher model to give pseudo labels, can a model become its own teacher for the training?

As our final contribution, we propose a novel UDA approach, called *AdaptOR*, for joint person pose estimation and instance segmentation. We first propose to extend the Mask R-CNN [He 2017] using *Disentangled Feature Normalization (DFN)* to train on two statistically different domains: natural images from *COCO*, and OR images from MVOR or TUM-OR [Belagiannis 2016]. DFN replaces every feature normalization layer in the feature extractor of the backbone model with two feature normalization layers: one for the source domain and another for the target domain. We propose to modify the loss function for our improved design of the backbone model. We pass the features of the source and the target domains separately to the downstream heads needed for separate weighing of the losses for the two domains.

Given a backbone model with the ability to train on two statistically distinct domains,

we propose to exploit explicit geometric constraints on the different augmentations of the unlabeled target domain images to generate accurate pseudo labels and to use these pseudo labels to train the model on high- and low-resolution OR images. The geometric constraints need to satisfy *transformation equivariant constraints* by transforming the model's predictions to observe the same geometric augmentations as of the input image. These explicit geometric constrains help the model to adapt to the target domain effectively.

Evaluation of the method on the two target domain datasets, *MVOR* and *TUM-OR*, with extensive ablation studies, show the effectiveness of our approach. The significantly better results on low-resolution images encourage the use of our method for the privacy-sensitive OR environment.

1.5 Outline

We first present related work in unsupervised domain adaptation and person localization methods in chapter 2. We then describe the unlabeled training and test datasets, evaluation metrics and performance of the state-of-the-art approaches in chapter 3. This Work has been published in [Srivastav 2018]. Chapter 4, published in [Srivastav 2019], puts forward the idea of two different image modalities as two distinct domains and describes our UDA approach to adapt the model to low-resolution OR depth images for the task of 2D HPE. Chapter 5 describes our UDA approach to train a model using the predictions coming from the multi-stage complex teacher model to train a generic model applied to low and high-resolution images from the target domain for joint 2D/3D pose estimation. This work has been published in [Srivastav 2020]. Finally, in chapter 6, we propose UDA approach for joint person pose estimation and instance segmentation that does not rely on complex multi-stage teacher network to provide the pseudo-labels. Instead, we propose a novel framework that exploit spatial and geometric constrains on the different augmentations of the unlabeled image to provide accurate pseudo labels. The work has been submitted to the journal Medical Image Analysis. Finally, some potential applications and the conclusion of the thesis with interesting new ideas for future work are presented in chapter 7 and chapter 8.

We also describe in Appendix A an iterative self-supervised approach for face detection in the OR, which has been published in [Issenhuth 2019]. This work won the runner-up award in the "Bench-to-Bedside category" at IPCAI 2019.

2 Related work

It is not the strongest of the species that survives, nor the most intelligent that survives. It is the one that is most adaptable to change.
– Charles Darwin

A cartoon illustration of the "street light effect". By drawing a similar comparison with the current deep learning literature, a model trained in one domain, when applied to a visually different target domain, results in a significant decrease in the performance. The domain adaptation aims to adapt an initial model trained on a source domain distribution to a different target domain distribution with few or no new annotations.

Chapter Summary

This chapter describes related literature on the domain adaptation methods applied to various computer vision tasks, ranging from image classification and object detection to semantic segmentation. Then, we discuss different computer vision approaches applied to low-resolution images to improve privacy. Finally, we describe various fine-grained person localization approaches ranging from 2D and 3D Human Pose Estimation (HPE) estimation to person instance segmentation.

2.1 Domain adaptation

2.1.1 Problem definition

Standard supervised learning approaches assume that the data, $\mathbf{d}$, and the corresponding labels, y, are drawn from a distribution, $\mathcal{D}$, during *training* to minimize some defined loss between the model's predictions, p, and the true labels, y. One of the key assumption being that the testing time data $\mathbf{d_{test}}$ will also be drawn from the same distribution $\mathcal{D}$. The performance guarantees of the model are measured on this assumption. In the domain adaptation, however, the assumption is that there exists a large labeled source domain dataset, $\{\mathbf{x}, y\}$ drawn from the distribution $\mathcal{X}$. At the test time we assume the model can receive the data, u, from a different domain $\mathcal{U}$. The goal of the domain adaptation is to learn to adapt the source domain model to the target domain.

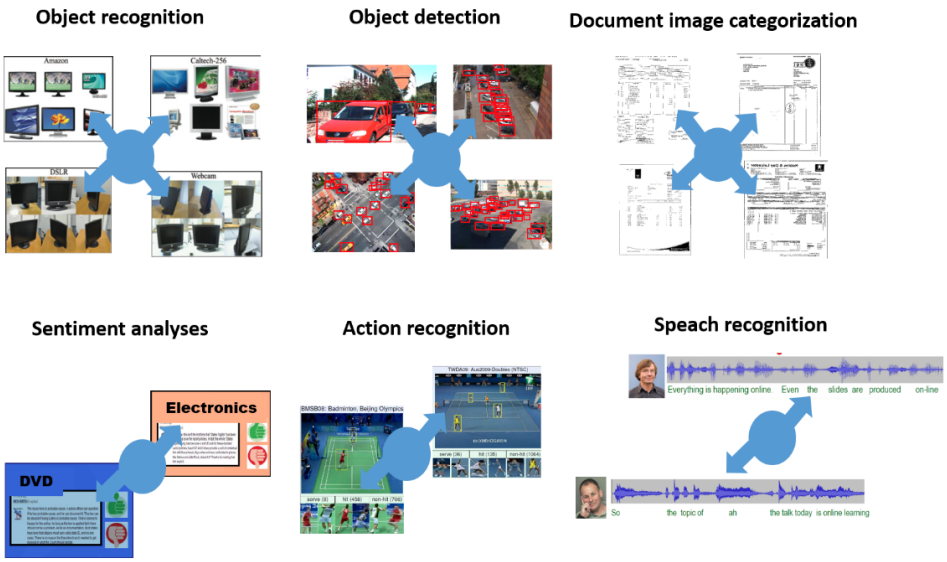


Figure 2.1: Different example cases for domain adaptation, image courtesy [Csurka 2017]

It can be achieved by a small set of labeled data from the target domain, not necessarily to tra in a model fully on the target domain data alone, but enough to adapt information from a model trained on a source domain to a target domain. These approaches are called *supervised domain adaptation* or *transfer learning*.
Conversely, *unsupervised domain adaptation* is the learning problem to adapt the model to the target domain, but without having access to any ground truth labels from the target domain. The main idea explored in the Unsupervised Domain Adaptation (UDA) is to learn the domain independent features to align the two distributions in some low-dimensional embedding space. In the next section, we provide a brief discussion to cover some related domain adaptation work in Automatic Speech Recognition (ASR) and Natural Language Processing (NLP) followed by a discussion on the visual domain adaptation.

2.1.2 Domain adaptation for ASR and NLP

The ASR approaches aim to convert the spoken language into text. The ASR has significantly improved over the past few years through improved neural networks and large-scale annotated datasets. Domain adaptation has started to become a key component in ASR, where the set of labeled human speeches are considered source domains and a particular person who uses the ASR system is considered the target domain. Approaches using teacher-student [Meng 2019] and self-training [Khurana 2021] learning have been used for the domain adaptation for ASR.
Domain adaptation has been extensively studied in the NLP, where the focus is to train a model on a large corpora of language text for a particular end task, such as sentiment analysis or document summarization, and then apply the learned model to the related but distinct domains. Several methods such as structural correspondence

learning [Liu 2010], feature replication [Daumé III 2009] have been proposed for the NLP tasks.

2.1.3 Visual domain adaptation

The focus of this dissertation is on visual domain adaptation. Contrary to domain adaptation for ASR and NLP, the visual data has the problem of containing ill-defined domains. Therefore, the standard practice assumes that each collected dataset belongs to a single visual domain while adaptation algorithms are evaluated across source and target domain datasets. The current UDA approaches for the different end tasks can be broadly classified in two main areas: *adversarial domain alignment* and *self-training*.

2.1.3.1 Adversarial domain alignment

The main idea in *adversarial domain alignment* based UDA approaches is to update either the feature, input, or output space from the target domain such that they are distributed in the same way as the source domain. At the feature space, for example, the domain invariant feature space is achieved using an additional neural network, called *domain classifier*, which essentially plays a min-max game with the *feature extractor* using adversarial learning [Goodfellow 2014]. Here, the *domain classifier* tries to fool the *feature extractor* by accurately distinguishing the source and the target domain features using a binary classification loss on the domain labels; the *feature extractor*, in turn, tries to fool the *domain classifier* by producing domain invariant feature such that the *domain classifier* would result in poor domain discrimination accuracy. The *adversarial domain alignment* at the feature space has been studied in [Ben-David 2010, Hoffman 2016, Chen 2018b, Chen 2019a, Du 2019, Saito 2019, Tran 2019, Hsu 2020, Sindagi 2020, VS 2021]

At the input space, these approaches utilize a *image-to-image translation* paradigm to mitigate the visual differences. The first method, Cycle-GAN [Zhu 2017], proposes to learn a function that can transform input image pixels across visual domains. These approaches, therefore, aim to transform the labeled source domain images to *target-like* source domain images. Since the annotations for the source domain images are readily available, the model trained on these *target-like* source domain images performs better on the original target domain images for a given end-task. These approaches aim to transfer the visual attributes - i.e., color, contrast - from the target domain to the source domain while preserving semantic content - i.e., shape and location of objects. The *adversarial domain alignment* at the input space has been studied in [Zhu 2017, Chen 2019c, Chen 2019d, Choi 2019, Li 2019]

Although these methods have made significant progress, stable training in the adversarial setup requires complicated training routines with careful adjustment of training parameters. Moreover, aligning the two domains using the *domain classifier* may not guarantee a required discriminative capability for a given end task.

2.1.3.2 Self-training

The *self-training* based UDA methods have emerged as promising alternatives to *adversarial domain alignment* as they follow a simple approach to learn the domain invariant representations. The main idea in the *self-training* is to generate pseudo labels on the unlabeled target domain by refining the predictions - generated from a given source domain trained model - using domain/task-specific heuristics, for example, confidence score in object detection [Deng 2021] or uncertainty in semantic segmentation [Liang 2019, Zheng 2021]. These pseudo labels are then used to train a model on the target domain jointly with the labeled source domain. The *self-training* has been extensively studied for object detection and semantic segmentation tasks [Inoue 2018, Zou 2018, RoyChowdhury 2019, Khodabandeh 2019, Kim 2019, Zou 2019, Zhao 2020a, Wang 2020a, Zheng 2021]. The *self-training* methods could further be improved in a *mean-teacher* framework to tackle noise in the pseudo labels [Cai 2019a, Liang 2019]. The main idea in the *mean-teacher* framework is to utilize two closely coupled models: a *teacher* model and a *student* model. The *teacher* model generates the predictions on the target domain unlabeled data, and the *student* model exploit these predictions for the training along with the source domain labeled data for training. The *student* model further improves the *teacher* model with the Exponential Moving Average (EMA). This mutual training improves both the *teacher* and *student* models gradually throughout the training. The *mean-teacher* and the *self-training* based UDA approaches have predominantly been inspired by the advances in the Semi-supervised Learning (SSL) [Tarvainen 2017, Berthelot 2019b, Sohn 2020a, Liu 2021b]. In fact, the UDA can be posed inside an SSL framework with the source domain data as the labeled and the target domain data as unlabeled along with additional complexity of the visual shift of the two domains. The *mean-teacher* paradigm has shown its effectiveness in the recent state-of-the-art approaches for self-supervised learning [Grill 2020, He 2020], semi-supervised learning [Sohn 2020a, Liu 2021b], as well as domain adaptation [Cai 2019a, Deng 2021].

2.1.3.3 Domain-specific feature learning

Some of the recent works aim to learn domain-specific feature representations instead of domain invariant using disentangled feature normalization. These approaches modify the feature normalization layers - as these control the feature distribution statistics - with two separate layers to disentangle the features from the two domains [Chang 2019]. The domain-specific features learning has been studied in the UDA for image classification [Chang 2019, Wang 2019], and federated learning on medical imaging [Li 2021]. It has also been used to boost performance in the supervised learning [Xie 2020], and adversarial robustness [Xie 2019]. The authors in [Wu 2021] comprehensively discuss the feature normalization under various visual recognition tasks. There also exist several survey papers that extensively discuss UDA for the end task of

image classification [Patel 2015, Wang 2018a, Zhuang 2020], semantic segmentation [Toldo 2020, Zhao 2020b], and object detection [Oza 2021].

A few notable works propose to use the UDA on the medical domain for cross-domain segmentation task [Li 2020a, Ouyang 2019, Orbes-Arteainst 2019, Chen 2019a], and image classification [Zhang 2020]. The authors in [Dong 2020] also study the UDA to identify domain invariant transferable features for endoscopic lesions segmentation. The authors in [DiPietro 2019] study the surgical workflow recognition with as few as one labeled sequence using SSL.

2.2 Low-resolution image recognition

The privacy-sensitive OR environment poses challenges in bringing the AI inside the OR. The recent controversies [Powles 2017] have raised public awareness regarding how personal data should be collected and controlled, along with how AI algorithms should use personal data in a privacy-safe way [Symons 2017]. One way to address these challenges is by using the federated learning [McMahan 2017] framework that allows training the model in a decentralized manner without explicitly sharing data. The federated learning has been recently used in medical imaging for segmenting the brain tumor [Sheller 2018] and detecting COVID-19 lung abnormalities in CT [Dou 2021]. Unlike medical imaging data, where privacy-sensitive information essentially lies in the metadata, direct video recording of OR using ceiling cameras contains the private information in the data itself. Adapting a model to very low-resolution images has been suggested in the literature to improve privacy [Chou 2018] that can further be incorporated inside the federated learning setup to improve multi-centric generalization. The low-resolution images entail significant degradation in the perceptual details of the image. It can be caused either due to poor image quality at the camera source or when capturing a large scene containing tiny objects. Low-resolution images can also be synthetically generated by applying handcrafted filters, for example, bicubic interpolation. These techniques, therefore, provide an automatic way to provide the training data to train a deep-learning model, called super-resolution (SR) models, that aims to learn a mapping function from low-resolution images to high-resolution images. The computer vision approaches enabling image recognition on low-resolution images for various end-tasks mainly utilize these super-resolution models to mitigate the spatial degradation of low-resolution images. Authors in [Bai 2018] employ Generative Adversarial Network (GAN) based architecture to directly regress for the high-quality SR face regions from blurry small input regions. The authors in [Shermeyer 2019] study the effect of various SR architectures for object detection in satellite imagery. They experimentally prove that the SR techniques enable the detection of tiny objects. The low-resolution image recognition has further been studied for 2D human pose estimation [Neumann 2018], face recognition [Ge 2018], image classification [Wang 2016], image retrieval [Tan 2018], object detection [Haris 2018, Li 2017], and activity recognition [Chou 2018, Ryoo 2017] tasks.

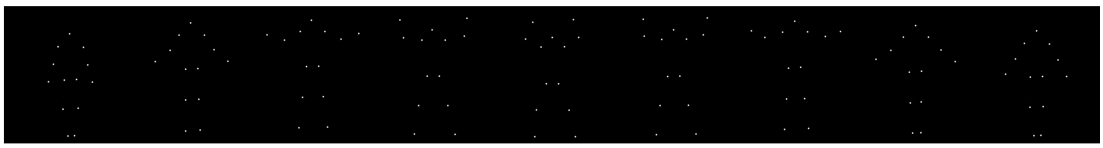


Figure 2.2: Authors in [Johansson 1973] showed in their seminal work that motion of few dots on the human body give rise to a compelling motion. See illustrative videos at[1]

In medical imaging, it has been mainly studied for image classification problems. Authors in [Chen 2021] present an integrated approach of super-resolution and medical diagnostic on the wireless capsule endoscopy and histopathological images. They first enhance the low-resolution images using a super-resolution framework and then use enhanced and low-resolution images in their proposed diagnosis classification network for accurate disease classification. Authors in [Mahapatra 2019] propose a progressive GAN and a triplet loss to improve the image quality that further improves the performance on vasculature segmentation and microaneurysm detection.

2.2.1 Privacy-preserving approaches using low-resolution images

As the low-resolution images significantly degrade the image quality, it can provide an effective means to develop privacy-preserving approaches [Haque 2017, Chou 2018, Gochoo 2020]. The authors in [Haque 2017, Chou 2018] propose to utilize low-resolution depth images for preserving privacy. They used an *off-the-shelf* super-resolution model to increase the perceptual details in the input depth image. The super-resolved image is then further passed to the action classification for the activity detection. The authors in [Gochoo 2020] propose to utilize ultra low-resolution thermal images indoor posture recognition. They propose to use a shallow architecture for the training on these privacy-preserving low-resolution images.

2.3 Person localization approaches

This section reviews various approaches for person localization, such as person instance segmentation and 2D/3D HPE on color and depth images. Person instance segmentation aims to estimate segmentation masks of each person in the image while differentiating each segmentation mask from the other one. HPE aims to localize different body parts, for example, eyes, nose, etc., either in 2D or 3D. Figure 1.4 shows example outputs for person localizations ranging from coarse person bounding box detection to fine-grained person instance segmentation and 2D/3D HPE.

2.3.1 Human pose estimation

HPE is a challenging problem as human poses are very high dimensional, showing different articulation and unusual motions. Moreover, there is an inherent occlusion

[1] https://youtu.be/1F5ICP9SYLU and https://youtu.be/rEVB6kW9p6k

problem due to the proximity of a person to other persons or other objects. There is also a loss of 3D in the 2D projection of images.

Almost half a century ago, Johansson et al. showed in their experiment that just a few dots (keypoints) on a human body could provide a compelling sense of human motion such as walking, running, and dancing [Johansson 1973], see figure 2.2. In the same year, Fischler et al. proposed the first computational approach called pictorial structure, where the idea was first to detect different parts in a bottom-up approach and then join them into a network-like structure [Fischler 1973]. These two seminal works provided a key influence and impelled the computer science community towards HPE. In the following, we discuss various HPE approaches that estimate keypoints either at 2D, on the image coordinates, or infer 3D orientation of the poses.

2.3.1.1 2D human pose estimation

Current 2D HPE approaches have been mainly studied either using bottom-up (keypoint-first) or top-down (person-first) approaches. The bottom-up approaches first detect all the keypoints for all the persons and then use a group post-processing method to associate keypoints to person instances; conversely, the top-down approaches first obtain the bounding box for each person instance using an *off-the-shelf* object detector and then employ a single-person pose estimation method to get the keypoints. The group post-processing methods in bottom-up approaches include Part Affinity Fields in CMU-Pose [Cao 2017], Part Association Field in PifPaf [Kreiss 2019], and Associative Embedding (AE) in [Newell 2017, Cheng 2020]. The leading methods for single-person pose estimation in the top-down approaches include Simple-Baseline [Xiao 2018], Alpha-Pose [Fang 2017], Cascaded-Pyramid-Network [Chen 2018a], HRNet [Sun 2019], and EvoPose2D [McNally 2020].

The bottom-up approaches are computationally faster due to their person-agnostic keypoint localization but yield inferior accuracy compared to the top-down approaches. The two-stage design in the top-down approaches helps them achieve significantly better accuracy, but at a more computational cost. Built on top of anchor-free detectors [Tian 2019b], some recent approaches such as DirectPose [Tian 2019a] and FCPose [Mao 2021] consider the keypoints as a special bounding-box with more than two corners and propose to regress the keypoint coordinates directly.

2.3.1.2 3D human pose estimation

3D HPE aims to predict locations of body joints in 3D space. This section focuses on the deep-learning based approaches that estimate 3D HPE from monocular RGB images either for a single-person or multi-persons.

3D single-person pose estimation

Single-person pose estimation approaches mainly use 3D datasets, such as Human3.6M [Ionescu 2013] or HumanEva [Sigal 2010], which provide both 2D and 3D

pose annotations, but with a single person performing actions in a controlled environment. Most approaches therefore leverage these datasets to learn an effective mapping from a given 2D pose to a 3D pose.

Authors in [Martinez 2017] designed a simple 2D-to-3D lifting network using few residual-based fully-connected layers and showed state-of-the-art performance on the Human3.6M dataset. A matching strategy is proposed [Chen 2017] to utilize 3D pose datasets of 2D to 3D mapping libraries and give estimated 2D pose. Authors in [Moreno-Noguer 2017] propose to encode pairwise distances of 2D and 3D body joints into two Euclidean distance matrices. They then train a regression network to learn the mapping of the two matrices. Authors in [Wang 2018b] propose to predict depth rankings of human keypoints as a viable cue for the 3D keypoint inference. Authors in [Yang 2018] adopted a 3D pose generator from the authors of [Zhou 2017] and propose a multi-source discriminator utilizing a given image, pairwise geometric structure, and joint location information.

3D multi-person pose estimation

3D multi-person pose estimation from a single image is an ill-defined problem because it involves inferring the relative positioning of the persons in the image in 3D from only a 2D image. It is a new field, and most approaches resolve the relative positioning by using various heuristics.

A bottom-up approach is proposed in [Mehta 2017] by using RTPose [Cao 2017] to infer person instances and further propose an Occlusion-robust Pose-maps (ORPM) to provide multi-style occlusion information irrespective of the number of people in an image. A top-down approach is proposed in [Rogez 2017] by using different models in a three-stage pipeline: faster R-CNN to detect person pose proposals, a classifier to score the different pose proposals, and a regressor network that refines pose proposals both in 2D and 3D. Authors in [Dabral 2019] used a similar top-down approach by modifying mask head in Mask R-CNN [He 2017] with an hourglass based 2D heat-map estimator and the 2D-to-3D lifting network from [Martinez 2017]. They further propose an optimization approach to determine the root-location in the 3D scene of a lifted root-relative 3D pose.

2.3.1.3 Human pose estimation in the OR

HPE in the OR is a relatively new field with approaches applied to either single or multi-view images and on color (RGB), depth (D), or both color and depth (RGB-D) images. The initial work [Kadkhodamohammadi 2014] propose a method to consistently track the upper body poses by offline optimization using discrete Markov Random Field (MRF) on the short RGB-D video sequences. The authors further propose an approach using the pictorial structure model [Fischler 1973, Felzenszwalb 2005] initially designed for the RGB images to the RGB-D images with a handcrafted Histogram of Depth Difference (HDD) features [Kadkhodamohammadi 2015]. Subsequent work use the multi-view RGB images [Belagiannis 2016] and multi-view RGB-D

images [Kadkhodamohammadi 2017c, Kadkhodamohammadi 2017a] for 3D HPE along with the corresponding multi-view RGB and multi-view RGB-D extensions to the pictorial structure model. Some recent work utilizes multi-view depth data for 3D HPE in the OR either using a voxel-based model [Hansen 2019] or point R-CNN model [Bekhtaoui 2020].

2.3.2 Person instance segmentation

Instance segmentation aims to identify the semantic class of each pixel as well as associate each pixel with a physical instance of an object. The task is challenging as it combines the two computer vision tasks into a single framework: object detection, where the goal is to classify individual objects and localize each using a bounding box and semantic segmentation, where the goal is to classify each pixel into a fixed set of categories without differentiating object instances. Similar to 2D HPE, the instance segmentation approaches can also be categorized into the bottom-up and top-down approaches. The top-down method also uses a two-stage design first to detect the bounding box and then either classify mask proposals or estimate segmentation masks from the bounding box proposals [He 2017, Chen 2019b, Bai 2017, Liu 2017, Lee 2020]. Similarly, the bottom-up methods associate pixel-level semantic segmentation output to the object instance. These approaches start from per-pixel classification results (e.g., FCN outputs), and attempt to cut the pixels of the same category into different instances [Zhang 2016, Liang 2017, Kirillov 2017, Arnab 2017]. Inside the OR, the only related work [Li 2020b] addresses a 3D scene semantic segmentation from multi-view depth images; however, the data is obtained from simulated clinical activities.

2.3.3 Joint person pose and instance Segmentation

A few notable works address the joint person pose estimation and instance segmentation [Papandreou 2018, He 2017, Zhang 2019b, Zhou 2020]. The authors in [Zhang 2019b, Zhou 2020] use pose estimation as a strong prior for the person instance segmentation. The PersonLab [Papandreou 2018] as a bottom-up method and Mask R-CNN [He 2017] as a top-down method are designed for the joint person pose estimation and instance segmentation.

2.4 Thesis positioning

Fine-grained person localization approaches that can not only exploit the abundant unlabeled and unseen data but also tackle Operating Room (OR) privacy are needed to develop and scale up novel assistance applications for the OR. In this chapter, we explore extensive literature in mainly three dimensions: domain adaptation, person localization architectures, and privacy-preservation using low-resolution. In the following, we highlight key points relevant to our proposed approaches described in this dissertation. As discussed in section 2.3.1, the multi-stage models for HPE provide state-of-the-art localization accuracy. In contrast, single-stage models are faster and therefore more

suited in the development of real-time applications. We have explored state-of-the-art two-stage models, for example, cascade Mask R-CNN [Cai 2019b] followed by HRNet [Sun 2019] as a teacher network to generate accurate pseudo labels. While as a student model, we extend run-time optimized models such as RTPose [Cao 2017], and Mask-RCNN [He 2017] for the real-time deployment in the OR. As discussed in section 2.3.1.2, a few residual-based fully-connected layers network to learn 2D to 3D pose mapping provide state-of-the-art accuracy on Human3.6 dataset [Ionescu 2013, Martinez 2017]. We, therefore, propose to extend Mask-RCNN by integrating *2D-to-3D* lifting network for end-to-end multi-person joint 2D/3D human pose estimation.

The section 2.2.1 explores how low-resolution images can provide possible directions towards building privacy-preserving approaches for the OR. Adapting the models to these spatially degraded low-resolution images is, however, challenging. Different from the current literature, which handles the low-resolution images by using *off-the-shelf* super-resolution models to enhance the spatial details, we directly adapt the features of the model for a given end-task by proposing two strategies. First, we propose integrating a feature-based super-resolution architecture in the end-to-end pipeline without generating intermediate super-resolution images. Second, we utilize advanced data augmentations to enforce consistency constraints between the high- and the low-resolution images derived from the pseudo labels, consequently enhancing the features for the low-resolution image.

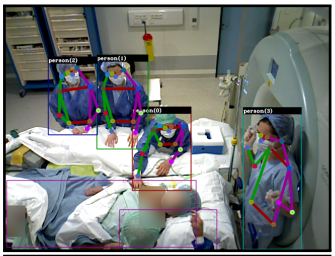
As discussed in section 2.1.3.2, the *self-training* based approaches provide a robust paradigm for visual domain adaptation. These approaches adapt the model to the target domain by training the model on its refined predictions. Training the model on these refined predictions, also called *pseudo-labels*, optimizes its feature for the given end task. Refining the predictions to generate accurate pseudo labels is however challenging. We have proposed two strategies for generating accurate pseudo-labels. First, we propose to exploit complex and multi-stage models to generate pseudo labels on the target domain to train a faster and single-stage model. Here the accurate pseudo labels are generated by utilizing *knowledge-* [Hinton 2015, Zhang 2019a] and *data-distillation* [Radosavovic 2018]. Second, we propose to exploit spatial and geometric constraints on the different augmentations of the unlabeled target domain image to generate accurate pseudo labels. The *mean-teacher* framework [Tarvainen 2017], as discussed in 2.1.3.2, helps to stabilize the training by handling noise in the pseudo labels.

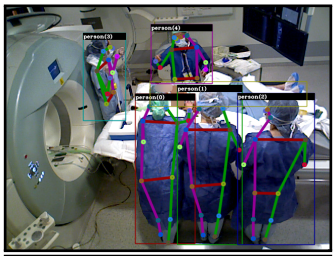
As discussed in section 2.1.3.3, we propose to extend the backbone model with disentangled feature normalization layers for simultaneous training on the source and target distributions.

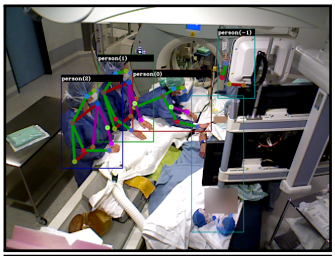
Part II Contributions

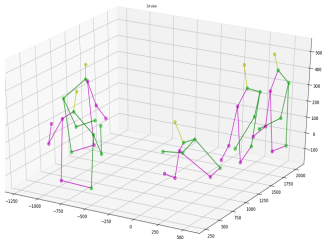
3 MVOR: Multi-view operating room dataset

You can have data without information, but you cannot have information without data.
– Daniel Keys Moran


View 1


View 2


View 3


3D View

Multi-view operating room (MVOR) is the first public dataset recorded during real clinical interventions. Dataset and code are available here: https://github.com/CAMMA-public/MVOR

Chapter Summary

3.1 Introduction

The availability of large-scale annotated datasets has been key in spurring interest and progress in human pose estimation and instance segmentation. 2D datasets captured in natural and *in the wild* environments such as MPII [Andriluka 2014], COCO [Lin 2014], and OCHuman [Zhang 2019b] include scenes with a wide amount of variability. 3D datasets, such as Human3.6M [Ionescu 2013], HumanEva [Sigal 2010], provide both 2D and 3D annotations, but with a single person performing actions in a controlled environment. As obtaining 3D ground truth on real-world images is an inherently difficult task, most approaches leverage these datasets to learn an effective mapping from 2D pose to 3D pose. The mapping is however learned from the simulated activities that do not cover real-world challenges. Therefore, models trained on such data do not generalize well to the challenging complex scenes such as OR. The TUM-OR dataset introduced in [Belagiannis 2016] is a multi-view OR dataset with 2D and 3D human poses. However, the dataset was captured during activities simulated by actors.

As our first contribution, we introduce the Multi-view Operating Room (MVOR) dataset, which is the first public Operating Room (OR) dataset captured during real surgical interventions. The MVOR dataset illustrates the complexity of a visually

different and challenging OR environment at a global and instance level. At a global level, it shows particular lighting conditions and instrument clutter in the OR. At the instance level, it shows clinicians wearing loose clothes and surgical masks and occluding one another due to close proximity. The MVOR dataset consist of 732 synchronized multi-view frames recorded by three RGB-D cameras in a hybrid OR. The dataset however is small and not suitable to train the deep learning algorithms and proposed to use as a test dataset to evaluate a model's ability to generalize to unseen configurations and color distribution. For training, we use an unlabeled dataset from the same OR, called *MVOR-unlabeled*, consisting of 80k synchronized color and depth images. In this chapter, we present the datasets, its ground-truth annotations, as well as baseline results from several approaches for person detection and 2D/3D human pose estimation.

3.2 MVOR

The MVOR dataset consists of RGB-D images sampled from eight days of recording in an interventional room at the University Hospital of Strasbourg during vertebroplasty and lung biopsy procedures. The images were captured using three synchronized RGB-D cameras (Asus Xtion Pro) mounted on the ceiling using articulated arms. The synchronized multiple cameras help to create a multi-view frame where the multi-view frame consists of RGB-D images recorded from all the cameras simultaneously. The cameras were mounted in such a way as to capture the key activities around the operating table, as shown in figure 3.1. The image and depth data were captured at 20 FPS in 640x480 VGA resolution using a recording software developed in-house. The intrinsic camera parameters of each camera were computed using a calibration pattern. The rigid transformation between the cameras and transformation of each camera to the global coordinate system were done in the two-step process described in [Svoboda 2005, Rodas 2015]. The operating table was considered to be the reference for the global coordinate system.

3.2.1 MVOR training set: *MVOR-unlabeled*

The unlabeled training set, called *MVOR-unlabeled*, consists of 80k RGB-D images sampled from four days of video recording. The full-day recordings of these videos include the frames when no activities are happening in the OR, so to select proper frames, we use OpenPose [Cao 2017], a multi-person pose estimator, on these videos to get the approximate number of persons in each frame. The computational efficiency of OpenPose allows us to make the inference on all the recording videos in a reasonable time. We divide the images into four categories: images with one, two, three, and four or more detected persons. Since OpenPose also gives a confidence score for each detected skeleton, we average the scores of the detected skeletons and take the 20k highest-scored images from each category (i.e., 80k images overall). This selection method ensures that the images contain persons in different numbers. We use the *MVOR-unlabeled* as the unlabeled training dataset to develop various unsupervised domain adaptation

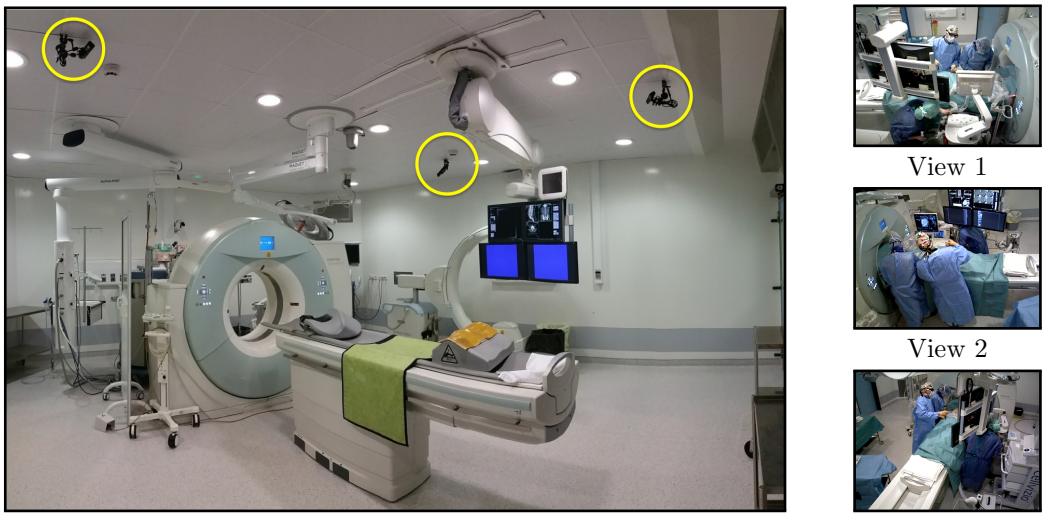


Figure 3.1: Multi-view setup and corresponding views in a room from the Interventional Radiology Department at the University Hospital of Strasbourg.

Figure 3.2: Illustration of the blurring process for the public release. The face of the patient, nudity and the eyes of the staff have been blurred.

approaches as explained in the following chapters.

3.2.2 MVOR test set: *MVOR* and *MVOR+*

3.2.2.1 Data

The test dataset of *MVOR* consists of 2196 frames sampled from four days of recording. The four days of videos are different from the ones of *MVOR-unlabeled* dataset to ensure the absence of overlap between the unlabeled training dataset and the test dataset. For public release of the dataset, the color images are needed to be blurred to ensure the

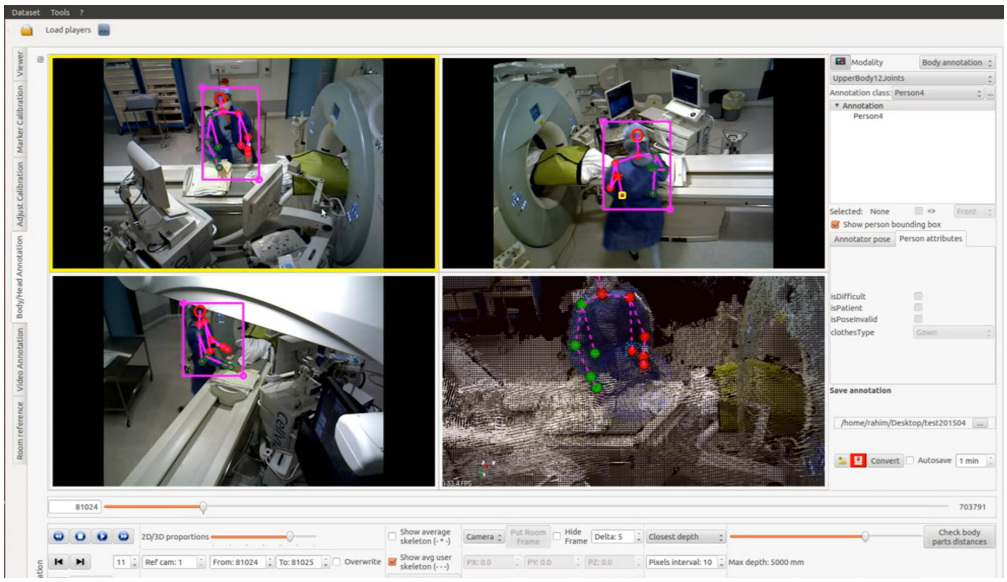


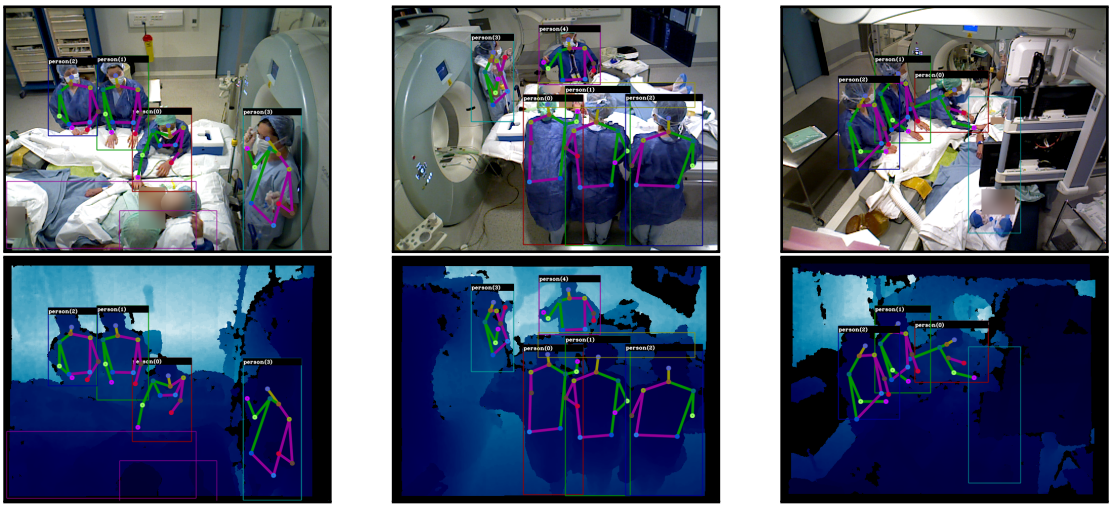
Figure 3.3: The tool used to generate the annotations, displaying the three views and the 3D point cloud in the interface. Right side body parts are shown in green and occluded body parts are marked by crosses. The annotators can move the joints in either 2D or 3D.

anonymization of the data. Patient faces and nude parts are fully blurred, while clinicians' faces are only blurred around the eyes when wearing a mask and fully blurred otherwise. A sample image is shown in figure 3.2.

3.2.2.2 Ground truth annotations

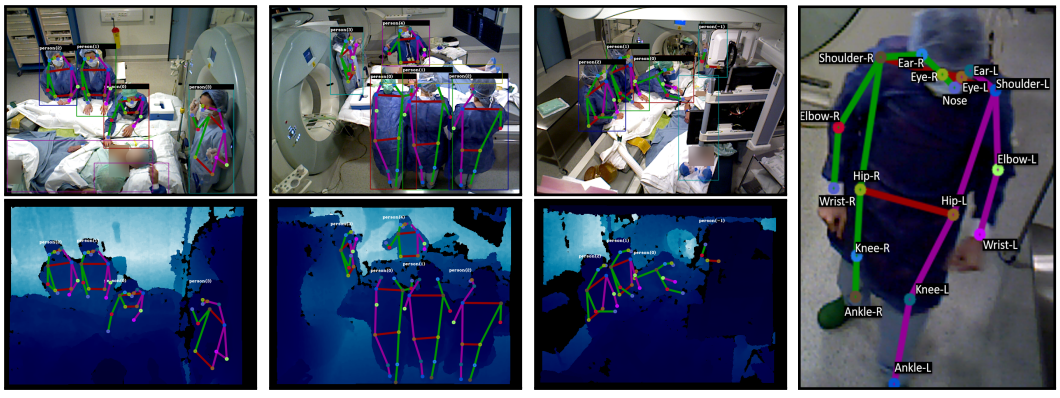
We release two iterations of the ground truth annotations for the MVOR test set: *MVOR* and *MVOR+*

- *MVOR*: It contains annotations for person bounding boxes and 2D/3D human poses. All persons are annotated with a full bounding box and staff who have more than 50% of their upper-body parts visible in at least one view are annotated with 2D and 3D upper-body pose keypoints. The 10 keypoints annotating the upper-body poses are shown in figure 3.4. To generate the annotations, we use a tool that displays all the three 2D views as well as the 3D point cloud, illustrated in figure 3.3. First, the annotator draws the 2D skeletons in all 2D views. To generate the 3D annotations, the 2D poses are back-projected into 3D using the depth information and initial 3D skeletons are computed by averaging all 3D skeletons across all views. We compute average 3D joint locations only among visible body joints. These initial 3D skeletons are not always accurate due to depth errors and differences in 2D joint annotations among the views, which are in turn caused by the large visual differences due to cameras rotation angles and partial occlusions. The annotator is therefore required to then ensure the quality of each 3D skeleton by verifying the accuracy of its projections to all views and by updating its locations directly in 3D when needed. Examples of available 2D/3D annotations are shown



View 1 View 2 View 3 annotation format 3D View

Figure 3.4: Visualization of 2D and 3D ground truth from the *MVOR* dataset



View 1 View 2 View 3 annotation format

Figure 3.5: Visualization of the ground truth from the extended *MVOR+* dataset

in figure 3.4.

- *MVOR+*: The *MVOR* dataset does not contain the annotations for all the persons, and 2D keypoints are only annotated for ten upper body parts. We extend the dataset with additional annotation to complete the annotations for all the

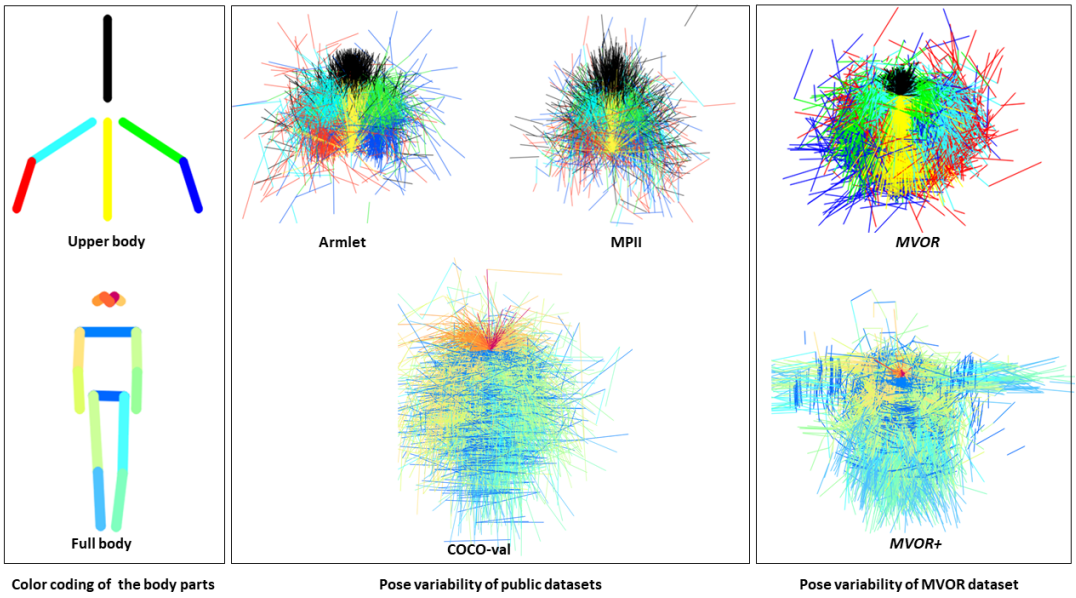


Figure 3.6: Visualization of 2D pose variability of upper-body poses from *MVOR* and full-body poses from *MVOR+* datasets. The 2D pose variability is also compared against the public Armlet [Gkioxari 2013], MPII [Andriluka 2014], and COCO [Lin 2014] datasets.

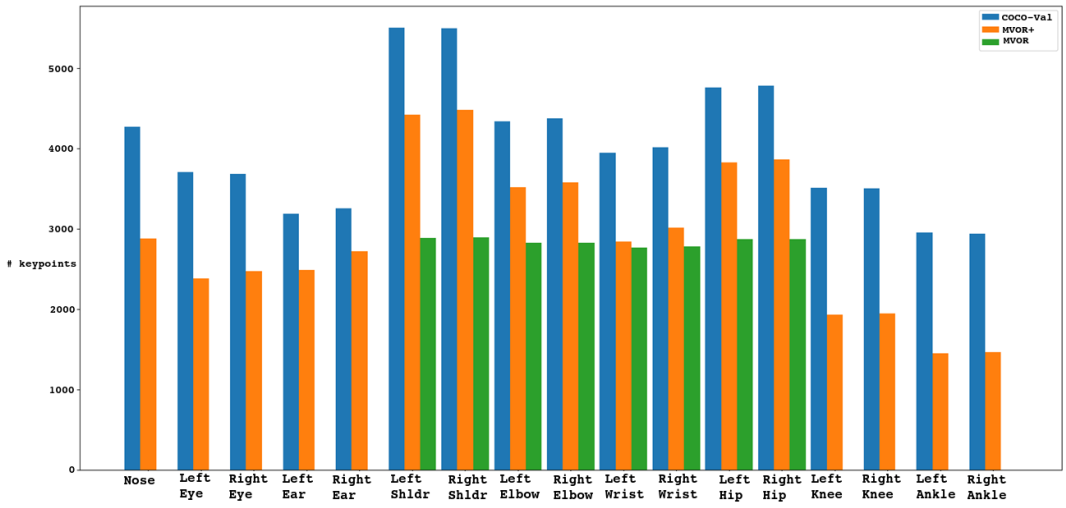


Figure 3.7: Statistics for the number of keypoints in the *MVOR*, *MVOR+* and COCO dataset. Our updated *MVOR+* dataset reaches close to the challenging COCO dataset in terms of the number of keypoints.

persons and use them in the standardized COCO evaluation framework. Before the extension, *MVOR* consists of 4699 person bounding boxes, 2926 2D upper body poses with ten keypoints, and 1061 3D upper body poses. The extended MVOR dataset, called *MVOR+*, consists of 5091 person bounding boxes and 5091 body poses with 17 keypoints in the COCO format; see example annotation in 3.5. We

use the visipedia tool[1] to extend the dataset in the COCO format. The *MVOR+* contains the same 3D annotations as *MVOR*.

Figure 3.6 shows the variability in upper-body and full-body poses from *MVOR* and *MVOR+*, respectively, along with a visual comparison against some of the natural image datasets. Figure 3.7 shows the statistics for the number of keypoints in the *MVOR* and updated *MVOR+* dataset and compares against the COCO dataset.

3.3 Comparison of state-of-the-art approaches

In this section, we present the comparison of state-of-the-art approaches for person detection and human pose estimation.

3.3.1 Compared person detection methods

Person detection can be performed using coarse bounding box detection or fine-grained instance segmentation. We evaluate several state-of-the-art approaches for person bounding box detection and person instance segmentation on *MVOR* and *MVOR+* datasets. These methods are trained on the large-scale *in the wild* COCO dataset for 80 object class categories. We evaluate these methods on the detections corresponding to only the person category. These approaches can be divided into two categories: one-stage and two-stage. The one-stage object detectors such as RetinaNet [Lin 2017b] for bounding box detection, and SoloV2 [Wang 2020c] for instance segmentation treat object detection as a regression problem on the fixed number of locations on the input image grid. These approaches take an input image and learn the class probabilities and bounding box coordinates. Two-stage object detectors such as Faster-RCNN [Ren 2015] for bounding box detection, and Mask-RCNN [He 2017] for instance segmentation first uses a Region Proposal Network (RPN) to generate bounding box proposals for the region of interests and then pass these proposals through separate heads for object classification and bounding-box regression and mask segmentation in the second stage. The two-stage detectors reach the highest accuracy rate than the one-stage detectors but are typically slower.

- RetinaNet [Lin 2017b]: This is a one-stage object detection approach. The authors hypothesize that the lower accuracy of the typical one-stage object detectors is primarily due to the extreme foreground-background class imbalance. They improve the performance of their one-stage object detector by proposing a novel focal loss that down-weights the loss assigned to well-classified predictions.

- SOLOv2 [Wang 2020c]: This is an improvement over the SOLO (segment objects by locations) [Wang 2020b]. SOLO is a single-shot approach for instance-segmentation that distinguishes different object instances based on the object centers and sizes. SOLOv2 further improves the SOLO by utilizing dynamic convolutions and matrix non-maximum suppression.

[1] https://github.com/visipedia/annotation_tools

- Faster-RCNN [Ren 2015]: This is a two-stage approach and the third object detector of the R-CNN family (R-CNN [Girshick 2014] and Fast-RCNN [Girshick 2015] are the first two). It enhances the R-CNN framework by making the region proposal network fully convolutional. Deep features are used instead of the input image to select the region of interests with a sliding window approach. Then, a second network classifies and refines the bounding box for each region of interest.

- Mask-RCNN [He 2017]: It extends the Faster-RCNN approach by adding an additional head for the object instance segmentation. The authors also proposed an ROI-align layer instead of ROI-pooling to improve the fine-grained pixel segmentation performance.

- Cascade-RCNN [Cai 2019b]: It improves the Mask-RCNN by proposing a sequence of multi-stage detectors that are trained with different Intersection over Union (IoU) thresholds.

The RetinaNet, Faster-RCNN, Mask-RCNN and Cascade-RCNN are evaluated using detectron2 framework [2] while SOLOv2 is evaluated using their official code [3].

3.3.2 Human pose estimation

We evaluate several approaches for 2D and 3D Human Pose Estimation (HPE) on the *MVOR* and *MVOR+* datasets. The 2D HPE has been mainly studied either using bottom-up (keypoint-first) or top-down (person-first) approaches. The bottom-up approaches first detect all the keypoints for all the persons and then use a group post-processing method to associate keypoints to person instances. Conversely, the top-down approaches first obtain the bounding box for each person instance using an *off-the-shelf* object detector and then employ a single-person pose estimation method to get the keypoints. The bottom-up approaches are computationally faster due to their person-agnostic keypoint localization but yield inferior accuracy compared to the top-down approaches. The two-stage design in the top-down approaches helps them achieve significantly better accuracy, but at a more computational cost. We evaluate RTPose [Cao 2017] as the bottom-up approach and Keypoint-RCNN [He 2017], Simple-Baseline (Simple-BL) [Xiao 2018], and HRNet [Sun 2019] as the top-down approaches. For 3D HPE, we assume the availability of 2D poses from 2D HPE methods and predict the 3D poses by lifting them from the 2D poses.

3.3.2.1 Compared 2D pose estimation methods

Bottom-up approaches

- RTPose [Cao 2017]: This is a bottom-up method, particularly well suited for real-time detections in RGB images. A deep multi-stage and two-branch CNN

[2] https://github.com/facebookresearch/detectron2
[3] https://github.com/aim-uofa/AdelaiDet/

jointly predicts heatmaps and part affinity fields to capture bodyparts and pairwise dependencies between body joints. Keypoints are then assembled into skeletons through a bipartite graph matching algorithm. We use the PyTorch version of RTPose in our evaluation[4].

- OpenPose [Cao 2017]: This is same as the RTPose but uses optimized inference models for faster inference. We use their official caffe version of OpenPose in our evaluation[5].

Top-down approaches

- Simple-BL [Xiao 2018]: The authors propose a simple single-person pose estimation model by adding a few deconvolutional layers on top of the ResNet backbone model. Despite its simplistic design, Simple-BL achieves competitive performance on the COCO dataset. We use the official PyTorch version of Simple-BL in our evaluation[6].

- HRNet [Sun 2019]: The authors propose a single-person pose estimation model that maintains the high-resolution feature map throughout the model design, helping it achieve much better localization accuracy and state-of-the-art performance. We use the official code of HRNet in our evaluation[7]

- Keypoint-RCNN [He 2017]: It is also a top-down approach extended from the Faster-RCNN by adding an additional head for the keypoint localization. However, unlike Simple-BL and HRNet that use an *off-the-shelf* object detectors for person bounding box detection, it uses shared features across the heads, making it computationally faster. We use the detectron2 API for the evaluation[8].

3.3.2.2 3D pose estimation

- 2D to 3D lifting from depth (*depth3D*): As the MVOR dataset consists of synchronized color and depth images, we lift the 2D poses to 3D from the corresponding depth image. However, depth images are usually noisy containing black patches on the image with zero depth value. To minimize the noise from the depth image, we estimate the depth value by calculating the median depth value inside the 15x15 bounding box around the 2D keypoint.

- 2D to 3D lifting using FCN (*2Dto3D*) [Martinez 2017]: The authors proposed a simple fully connected neural network to lift the 2D keypoints to 3D. The authors learn the mapping from the 2D/3D ground truth obtained from the Human3.6 dataset [Ionescu 2013]. The authors show that given the accurate 2D keypoint

[4] https://github.com/tensorboy/pytorch_Realtime_Multi-Person_Pose_Estimation
[5] https://github.com/CMU-Perceptual-Computing-Lab/openpose
[6] https://github.com/microsoft/human-pose-estimation.pytorch
[7] https://github.com/HRNet/deep-high-resolution-net.pytorch
[8] https://github.com/facebookresearch/detectron2

detections, their lifting network could accurately lift them to the 3D keypoints. We use the PyTorch version of *2Dto3D* in our evaluation[9].

3.3.3 Evaluation metrics

3.3.3.1 Person detection

We use the Average Precision (AP) $AP_{0.5:0.95}$ metric from COCO [Lin 2014] for the evaluation. The bounding box evaluation metric AP^{bb}_{person} uses IoU over boxes. As the *MVOR* and *MVOR+* datasets do not have a ground-truth for the person segmentation masks, we opt for an alternate approach for the instance segmentation evaluation. We evaluate the instance segmentation by computing a tight bounding box on the prediction masks and comparing them with ground-truth bounding boxes called $AP^{bb(from\ mask)}_{person}$.

3.3.3.2 2D human pose estimation

We use the AP $AP_{0.5:0.95}$ metric from COCO [Lin 2014] for the evaluation on the *MVOR+* dataset. The pose estimation evaluation metric AP^{kp}_{person} uses the Object Keypoint Similarity (OKS) over person keypoints to compare the ground and the predictions. As the *MVOR* dataset contains 10 keypoints in non COCO format, we use the mean percentage of correct keypoints (*meanPCK*) [Yang 2012] to compare the baseline pose estimation methods. This metric measures the localization accuracy of the body joints, based on the scale of the person. To match detected and ground-truth skeletons, a tight bounding box is computed for each ground-truth skeleton from its keypoints. Then, for each ground-truth skeleton, we select the detection with the highest confidence score among the detections which have more than 30% of their keypoints in the ground-truth bounding-box.

3.3.3.3 3D human pose estimation

We use the 3D Mean Per Joint Position Error (MPJPE) in millimeters (mm) to evaluate the 3D keypoints. The MPJPE metric is computed using the eight keypoints from the upper-body pose (shoulder, elbow, hand, hip), as both *MVOR* and *MVOR+* contain the same number of 3D poses with eight common 3D keypoints. The 3D ground-truth keypoints are expressed in the camera-coordinate frame, and each joint in the 3D pose is subtracted from the pelvis root-joint (taken as the mean of left and right hips) to obtain the root-relative pose. The root-relative pose is computed before calculating the MPJPE error.

High values are desired for the bounding box detection metric (AP^{bb}_{person}), person instance segmentation metric ($AP^{bb(from\ mask)}_{person}$), and 2D HPE metrics (*meanPCK* and AP^{kp}_{person}), while a low value is desired for the 3D HPE metric (MPJPE).

[9] https://github.com/weigq/3d_pose_baseline_pytorch

Table 3.1: Person bounding box detection and instance segmentation results from the state-of-the-art methods on *MVOR*, *MVOR+*. All these methods are trained on the large-scale annotated *COCO* dataset and evaluated on the *MVOR* and *MVOR+* datasets without any OR training. We also show the results on the *COCO* dataset for comparative analysis. The two-stage detectors perform better than the one-stage detectors. Increasing the model complexity also contribute to the increase in the accuracy. R50-FPN and R101-FPN correspond to the ResNet backbone with 50 and 101 layers, respectively, along with the Feature Pyramid Network (FPN). X101, X152 correspond to the ResNext backbone with 101 and 152 layers, respectively. X152-FPN-DConv is a very deep network that uses Deformable Convolution (DConv), particularly suited for object detection networks. It is trained for a much longer duration, helping it achieve better results. The significantly poor results on the depth (D) images and very low-resolution images (downsampled with 12x scale) are understandable as these images are not represented in the training dataset. The AP^{mask}_{person} results show the instance segmentation results on the COCO dataset by using ground truth person masks.

Models	***MVOR***				***MVOR+***				***COCO***	
	$\mathbf{AP^{bb}_{person}}$									
	RGB		D		RGB		D		RGB	
	1x	12x	1x	12x	1x	12x	1x	12x		
one-stage										
RetinaNet(R50-FPN)	47.10	23.37	12.40	03.25	46.30	22.61	11.81	03.12	52.52	
RetinaNet(R101-FPN)	48.60	23.17	13.70	04.70	47.79	22.46	13.08	04.59	53.43	
two-stage										
Faster-RCNN(R50-FPN)	48.10	24.69	15.30	07.65	47.31	23.88	14.58	07.25	54.47	
Faster-RCNN(R101-FPN)	49.70	22.88	17.60	05.90	48.85	22.04	16.85	05.62	55.67	
Faster-RCNN(X101-FPN)	50.30	20.63	14.80	06.40	49.38	19.88	14.24	06.10	56.58	
Mask-RCNN(R50-FPN)	49.10	25.77	13.70	05.63	48.50	24.90	13.15	05.35	55.30	
Mask-RCNN(R101-FPN)	50.20	25.19	16.00	05.98	49.31	24.27	15.29	05.73	56.56	
Mask-RCNN(X101-FPN)	50.80	19.81	13.90	04.61	49.91	19.22	13.45	04.33	57.65	
Cascade-Mask-RCNN(R50-FPN)	51.40	27.16	14.10	08.38	50.49	26.31	13.43	07.95	58.83	
Cascade-Mask-RCNN(X152-FPN-DConv)	54.50	25.20	11.50	03.53	53.88	24.38	10.88	03.39	62.21	
	$\mathbf{AP^{bb(from\ mask)}_{person}}$									AP^{mask}_{person}
one-stage										
SOLOv2(R50-FPN)	47.40	23.93	11.90	07.75	45.86	23.00	11.19	07.34	50.94	*45.87*
SOLOv2(R101-FPN)	49.39	21.12	16.43	01.95	47.87	20.39	15.67	01.87	52.36	*47.12*
two-stage										
Mask-RCNN(R50-FPN)	48.34	25.11	13.57	05.92	46.88	24.03	12.98	05.59	53.42	*47.66*
Mask-RCNN(R101-FPN)	48.74	24.49	15.90	06.05	47.20	23.31	15.05	05.71	54.25	*48.66*
Mask-RCNN(X101-FPN)	50.08	19.24	13.73	04.53	48.36	18.48	13.19	04.28	55.53	*49.69*
Cascade-RCNN(R50-FPN)	49.71	26.04	13.56	07.96	48.19	24.98	12.89	07.46	55.69	*48.58*
Cascade-RCNN(X152-FPN-DConv)	53.71	23.90	11.62	03.57	52.37	22.80	10.93	03.39	60.28	*52.41*

3.3.4 Results

3.3.4.1 Person detection

Table 3.1 shows the results for person bounding box detection and instance segmentation results on *MVOR*, *MVOR+* at original (1x: 640x480) and downsampled (12x: 53x40) images. The evaluated methods are trained on the large-scale annotated *COCO* dataset without any training on the OR images. We observe a similar trend from to natural images where the two-stage detectors perform better than the one-stage detectors. Increasing the model complexity also contributes to the increase in the accuracy. We observe a significant drop in the performance on the original resolution and much poorer results on the depth (D) images and very low-resolution images. The significantly poor results on the depth and the low-resolution images are understandable as these images are not represented in the training dataset. Figure 3.8 shows the qualitative results for person bounding box detection and instance segmentation from the state-of-the-art approaches.

3.3.4.2 Human pose estimation

Table 3.2 and 3.3 show the results for 2D HPE and 3D HPE on the *MVOR*, *MVOR+* and *COCO* datasets and 3D HPE on the *MVOR+* dataset at original (1x: 640x480) and downsampled (12x: 53x40) images. We evaluate these methods for different backbones and different input resolutions. The top-down approaches perform better than the bottom-up approaches. 3D HPE using the *2Dto3D* lifting network [Martinez 2017] performs better than lifting the poses from the corresponding depth images.

3.4 Conclusion

In this chapter, we present a new multi-view dataset for multi-person detection and 2D/3D human pose estimation in a challenging environment, namely a modern operating room, which contains inherent visual challenges such as multiple occlusions. We also present the results of several recent baseline methods. This dataset can thus be helpful to evaluate a detector's ability to generalize to unseen configurations and color distribution and assess the performance of 3D multi-person pose estimation methods on real-world data. We observe a decrease in the accuracy on the original scale (1x) and significantly poorer accuracy on the downsampled images (12x) and depth images. As discussed in chapter 1, the low-resolution images could effectively tackle the OR privacy; the following chapter therefore designs unsupervised domain adaptation approaches that work particularly well on the depth and low-resolution images.

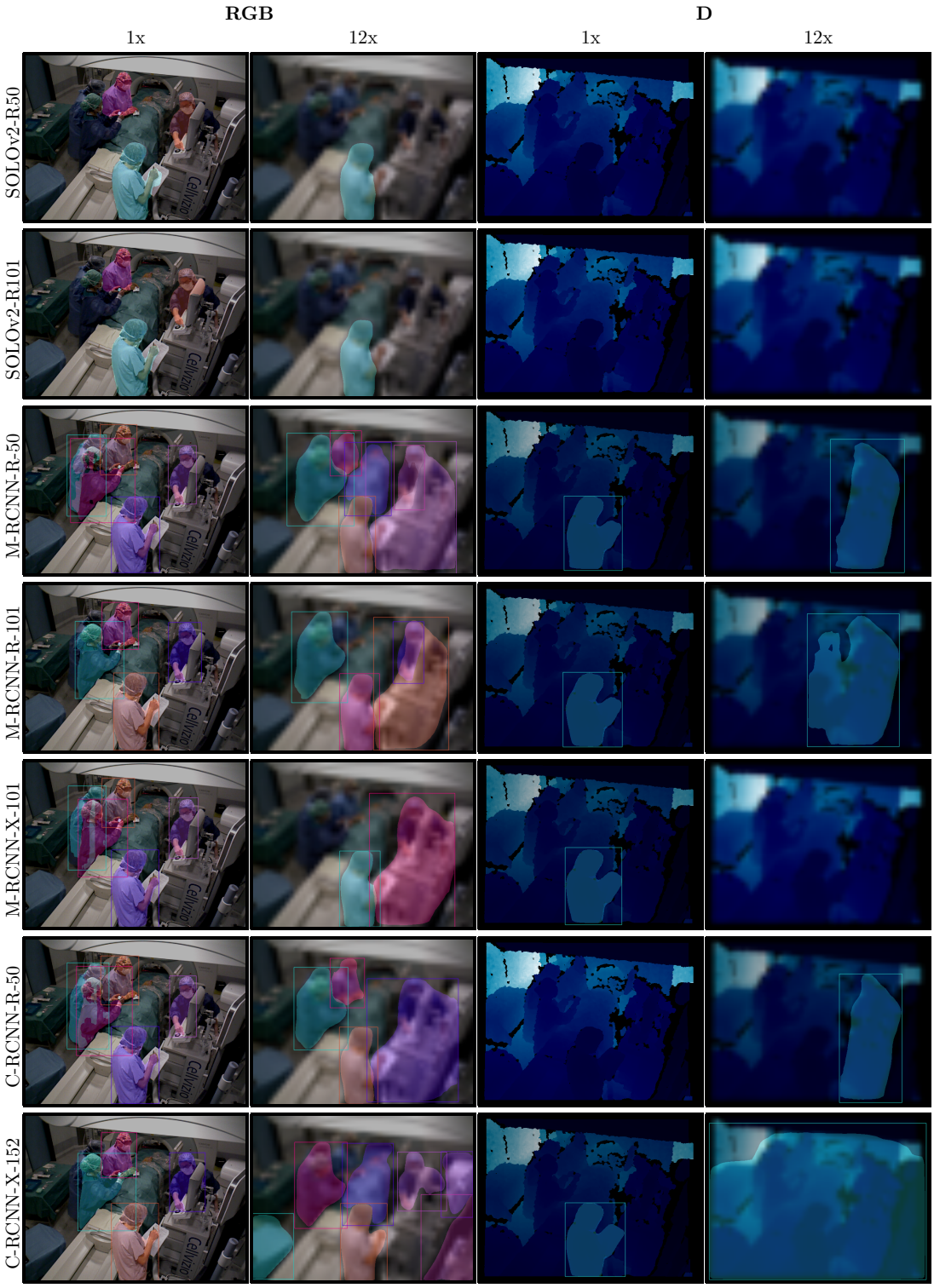


Figure 3.8: Qualitative results for person detection from the state-of-the-art approaches on a sample color and depth image from the *MVOR* dataset with downsampling factor 1x and 12x. M-RCNN-R-50, M-RCNN-R-101, M-RCNN-X-101, C-RCNN-R-50, and C-RCNN-X-152 correspond to Mask-RCNN(R50-FPN), Mask-RCNN(R101-FPN), Mask-RCNN(X101-FPN), Cascade-RCNN(R50-FPN), and Cascade-RCNN(X152-FPN-DConv), respectively.

Table 3.2: Results for 2D HPE on *MVOR*, *MVOR+* and *COCO* datasets and 3D HPE on *MVOR+* dataset at original resolution (1x: 640x480) for bottom-up and top-down approaches. Keypoint-RCNN is evaluated with three backbone networks ResNet with 50 and 101 layers and ResNext with 101 layers. The Simple-BL method is also evaluated with a ResNet backbone consisting of 50, 101, and 152 layers. The HRNet model is evaluated with W32 and a more deeper W48 network. Both Simple-BL and HRNet models are evaluated at the input resolution of 256x192 and 384x288 pixels. We use Mean Percentage of Correct Keypoints (meanPCK) on the *MVOR* dataset and AP^{kp}_{person} on the *MVOR+* dataset for 2D HPE. 3D HPE is evaluated by lifting the 2D coordinates to 3D from the corresponding depth images and by using a *2Dto3D* lifting network [Martinez 2017].

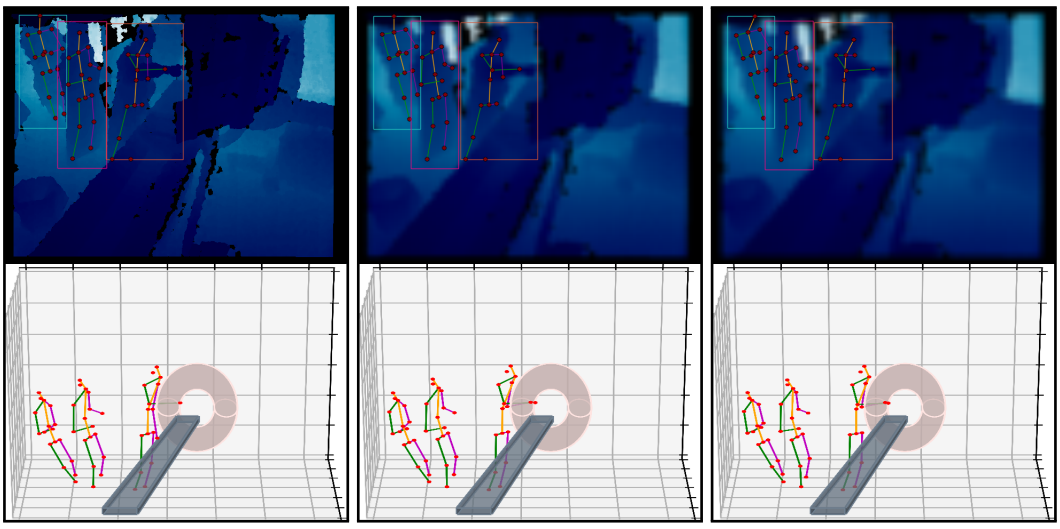
	2D Pose Estimation					**3D Pose Estimation**			
	MVOR		***MVOR+***		***COCO***	***MVOR+***			
	meanPCK		**AP^{kp}_{person}**			**MPJPE (in mm)**			
3D lifter →						*2Dto3D*		*depth3D*	
2D pose models ↓	RGB	D	RGB	D	RGB	RGB	D	RGB	D
bottom-up									
OpenPose	56.60	07.37	43.38	02.80	52.33	432.37	628.70	358.04	434.26
RTPose	70.28	10.55	33.38	01.25	46.21	522.75	775.85	413.12	694.23
top-down									
Keypoint-RCNN(R50-FPN)	74.70	17.55	45.67	04.84	65.50	144.15	319.45	264.61	443.28
Keypoint-RCNN(R101-FPN)	75.40	20.07	46.24	06.90	66.10	143.44	320.34	264.83	377.80
Keypoint-RCNN(X101-FPN)	74.30	20.43	46.17	06.12	66.00	145.18	340.47	266.16	389.10
Simple-BL(R50_256x192)	75.20	22.90	51.91	11.11	70.40	136.64	222.92	262.39	267.93
Simple-BL(R50_384x288)	74.90	22.30	52.46	08.83	72.20	139.62	232.30	264.74	268.05
Simple-BL(R101_256x192)	75.20	22.70	53.30	10.72	71.40	137.18	223.54	265.28	271.95
Simple-BL(R101_384x288)	75.30	23.10	54.12	10.76	73.60	137.37	221.39	265.80	265.31
Simple-BL(R152_256x192)	75.70	22.30	53.95	09.82	72.00	134.68	230.25	262.87	271.87
Simple-BL(R152_384x288)	75.50	23.80	54.59	12.01	74.30	135.81	218.23	261.67	262.44
HRNet(W32_256x192)	75.80	22.70	55.73	10.56	74.40	135.41	223.95	262.05	274.85
HRNet(W32_384x288)	76.40	22.00	56.39	08.98	75.80	133.84	230.49	260.94	268.66
HRNet(W48_256x192)	75.80	22.30	56.39	10.24	75.10	134.55	225.33	262.04	270.08
HRNet(W48_384x288)	76.00	20.40	57.10	07.30	76.30	134.54	234.94	260.82	276.44

Table 3.3: Results for 2D HPE on *MVOR*, *MVOR+* and *COCO* datasets and 3D HPE on *MVOR+* dataset at downsampled resolution (12x: 53x40). Images are upsampled to the original size after the downsampling before being fed to the models. We see significantly poor results especially on the AP metric from all the approaches on these heavily downsampled images.

	2D Pose Estimation				**3D Pose Estimation**			
	MVOR		***MVOR+***		***MVOR+***			
	meanPCK		**AP^{kp}_{person}**		**MPJPE in mm**			
3D lifter →					*2Dto3D*		*depth3D*	
2D pose models ↓	RGB	D	RGB	D	RGB	D	RGB	D
bottom-up								
OpenPose	36.55	08.10	12.43	00.64	567.04	714.97	450.12	559.74
RTPose	39.97	08.91	05.20	00.35	830.65	807.36	596.00	653.64
top-down								
Keypoint-RCNN(R50-FPN)	49.90	18.66	14.42	02.35	194.35	350.95	293.62	424.13
Keypoint-RCNN(R101-FPN)	51.90	16.37	15.53	03.55	192.39	325.75	295.86	379.71
Keypoint-RCNN(X101-FPN)	49.79	06.99	13.75	00.42	202.96	500.48	293.64	495.58
Simple-BL(R50_256x192)	62.86	18.48	25.53	07.35	166.67	238.92	265.95	293.65
Simple-BL(R50_384x288)	61.24	17.99	23.77	05.75	172.96	250.69	270.83	299.66
Simple-BL(R101_256x192)	63.19	16.80	26.84	06.19	163.12	250.66	264.48	309.81
Simple-BL(R101_384x288)	61.28	15.95	23.84	04.73	169.65	259.01	274.89	321.50
Simple-BL(R152_256x192)	63.88	18.15	26.35	06.21	162.37	245.55	267.83	302.68
Simple-BL(R152_384x288)	61.11	18.45	23.89	07.13	169.72	244.16	266.15	294.95
HRNet(W32_256x192)	53.59	15.99	19.94	05.73	193.60	280.91	310.96	352.02
HRNet(W32_384x288)	50.87	15.14	17.93	05.28	195.86	297.16	320.80	372.25
HRNet(W48_256x192)	54.57	15.87	21.32	06.02	186.68	286.57	312.26	352.35
HRNet(W48_384x288)	47.58	15.87	16.42	05.45	211.60	276.06	356.02	353.84

4 Domain adaptation across visual modalities for human pose estimation on low-resolution depth images

Supervision is the opium of the AI researcher.
– Jitendra Malik



1x 8x 10x

A sample qualitative result from our unsupervised domain adaptation approach on different low-resolution depth images for 2D/3D human pose estimation. Demo video is available here: https://cutt.ly/depthpose. Project page: https://github.com/CAMMA-public/ORPose-depth

Chapter Summary

4.1 Introduction

Color (RGB) and depth (D) images are different but offer complementary information. The RGB images contain rich texture details and are pervasive due to cheap visual sensors. Conversely, depth images are texture-less and encode the object's distance from the camera center. The current progress in the RGB-D visual sensors capturing synchronized color and depth images can open up the novel ways to develop automated assistance applications for the Operating Room (OR). Deploying these applications inside the OR environment is, however, challenging due to the use of these privacy-intrusive visual sensors. Direct processing of OR images at high-resolution can intrude the privacy, significantly in the RGB images, but also in texture-less depth images [Cheng 2017, Chou 2018]. This is particularly relevant in environments where the number of persons is limited and where the persons could potentially be more easily identified. As outlined in section 1.2.1 and illustrated in figure 4.1, the low-resolution images significantly degrade the spatial details. Therefore, these could provide a viable means to tackle OR privacy and to develop more privacy-compliant computer-vision applications inside the clinical institutions. In [Chou 2018], it has been shown that activity recognition can be performed on low-resolution depth images captured for the tasks of hand-hygiene classification and Intensive Care Unit (ICU) activity logging. In this chapter, as our first contribution, we first investigate whether **low-resolution depth images** contain sufficient information for accurate 2D/3D Human Pose Estimation (HPE). The large-scale RGB datasets for HPE such as COCO [Lin 2014] and MPII [Andriluka 2014] have recently shown remarkable progress on *in the wild* natural

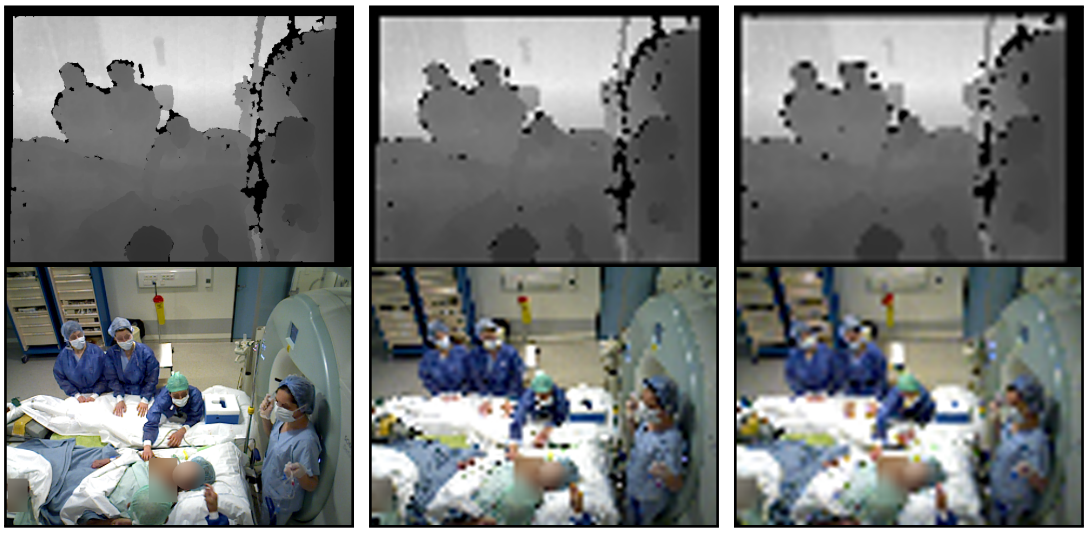
images. The availability of annotated depth data is however limited to either synthetic annotations [Shotton 2013] or real datasets [Haque 2016] recorded by the actors performing simulated actions. These annotations may not capture the intricacies of the challenging real-world environment such as OR.

Given the well performing 2D HPE models on the RGB images, as our second contribution, we proposed an Unsupervised Domain Adaptation (UDA) approach for 2D/3D HPE across visual modalities i.e from RGB to depth images. Our approach does not require any manual labels and only needs synchronized color and depth images during the training. We use *MVOR-unlabeled* as described in section 3.2.1 for training and propose to use the detections from a state-of-the-art method applied to the color images as pseudo labels for the corresponding depth images. This simple idea turns out to be very effective. Indeed, as our approach only requires a set of RGB-D images at train time, it can be easily retrained in any facility since no annotation process is needed. Then, it can run round the clock on low-resolution depth images from the same facility. We propose two strategies for effective low-resolution feature learning to tackle the loss of spatial details in low-resolution images. As our first strategy, we propose to integrate super-resolution feature maps in the bottom-up RTPose [Cao 2017] method that utilizes intermediate super-resolution feature maps to learn the high-frequency features better. As our second strategy, we exploit advanced data-augmentations such as low-resolution down- and up-sampling, rand-augment [Cubuk 2020] and random cut-out [DeVries 2017]. We use top-down Keypoint-RCNN [He 2017] model to train on these heavily augmented images along with the pseudo labels. We show significantly better results specifically on the low-resolution depth images for both of our strategies. The top-down Keypoint-RCNN model utilizing state-of-the-art data augmentations performs significantly better than the bottom-up approach. For 3D HPE, we investigate the lifting of 2D keypoints to 3D from the depth value and the use of a *2Dto3D* lifting network as explained in section 3.3.2.2. We observe better results from the *2Dto3D* network compared to naively getting the 3D depth value from the depth image.

4.2 Methodology

4.2.1 Pseudo label generation

In the literature, authors have either used manually annotated or synthetically generated datasets to train for HPE on depth images. Manual annotations can be expensive and time-consuming, and synthetic annotations are difficult to generate due to the constraint of realistic rendering and do not always generalize well to real scenarios. Therefore, we use an alternate approach to generate annotations. This approach is based on the observation that the RGB-D cameras capture synchronized color and depth streams, and recent HPE methods trained on the COCO dataset [Lin 2014] work remarkably well on the color images. Therefore, we use detections from the color images to train the model for the depth images. To facilitate this approach, we use *MVOR-unlabeled* dataset containing 80k synchronized color and depth images captured



1x(640x480) 8x(80x60) 12x(64x48)

Figure 4.1: Depth and color images from MVOR down-sampled at different resolutions using bicubic interpolation (resized for better visualization). Low-resolution depth images contain little information for the identification of patient and health professionals. Corresponding color images in the second row are shown for better appreciation of the downsampling process.

in the OR during real surgical procedures. Then, we use the state-of-art person detector Mask-RCNN [He 2017] and a single person pose estimator Simple-BL [Xiao 2018] on color images to generate detections. We filter out the false positives and retain high-quality detections in both the stages using thresholds selected from the qualitative results on a small set of images. The pseudo labeling process is shown in figure 4.2 (a). This approach generates pseudo labels truth automatically without using any human annotation efforts. It is therefore scalable and can be deployed to any facility. For human pose estimation, we choose here a two steps method based on Mask-RCNN and Simple-BL for their state-of-the-art performance on the public COCO dataset as shown in table 3.2.

4.2.2 Proposed architectures

We propose two strategies to learn better features for the low-resolution images and effectively exploit the pseudo labels in training the model. As our first strategy, we use the bottom-up RTPose [Cao 2017] model and extend it with super-resolution feature maps to better learn the high-frequency features. As our second strategy, we use top-down Keypoint-RCNN [He 2017] model and propose to exploit advanced data-augmentations to adapt the model to low-resolution images.

4.2.2.1 Bottom-up: ORPose-Depth(RT)

Our proposed model is inspired by the recent developments in the area of super-resolution and multi-person human pose estimation. We propose to integrate a super-resolution image estimator and a 2D multi-person pose estimator in a joint

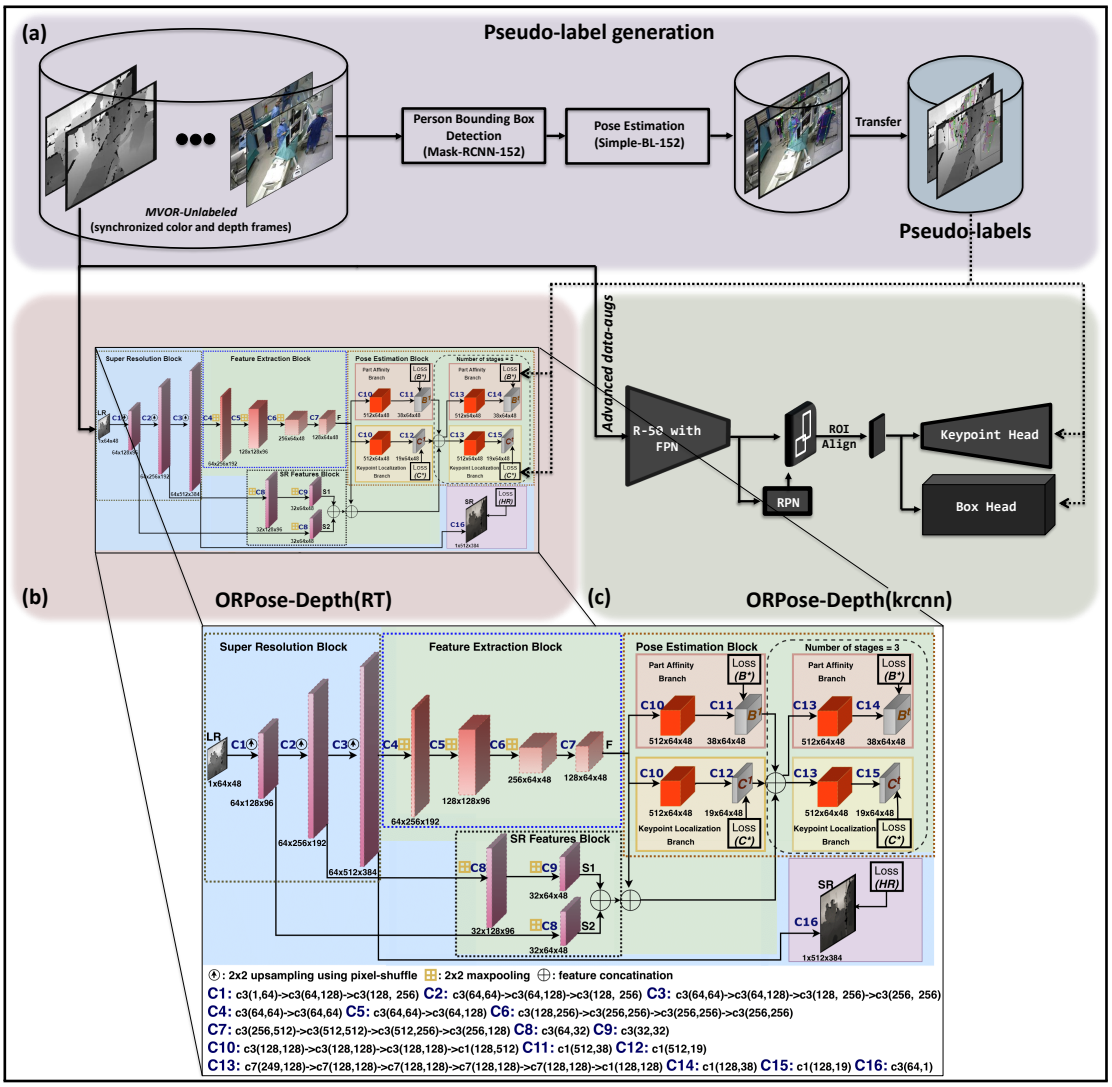


Figure 4.2: Proposed approach. (a) Pseudo-label generation: we use a person-bounding box detector (Mask-RCNN with ResNet-152) followed by a single person pose estimator (Simple-BL with ResNet-152) to generate the pseudo labels on the color images of *MVOR-unlabeled*. These labels are then transferred to the corresponding depth images. (b) ORPose-Depth(RT): we propose modifying bottom-up RTPose architecture to train the model on low-resolution depth images. The super-resolution block increases the spatial resolution by a factor of 8x and generates intermediate SR feature maps (S1, S2) used by the pose estimation block to learning high-frequency features. All losses are mean square error losses. C1 to C16 are convolution layers grouped together for better visualization and described below the figure, where c1(n1,n2), c3(n1,n2), c7(n1,n2) each represent a convolution layer with kernel size 1x1, 3x3, 7x7 and padding 0, 1, 3, respectively. Parameters n1 and n2 are the numbers of input and output channels, and all convolution layers are followed by RELU non-linearity. (c) ORPose-Depth(krcnn): we propose to utilize advanced data augmentations to better learn the low-resolution features in the top-down Keypoint-RCNN model

architecture, illustrated in figure 4.2 (b). This architecture is based on modification from the RTPose network [Cao 2017]. Besides yielding competitive results on COCO and

MVOR, RTPose has the advantage to perform multi-person pose estimation in a single step, thereby simplifying the integration and training of the super-resolution modules. It is composed of a *feature extraction block* and a *pose estimation block* shown in figure 4.2 (b).

We introduce a *super-resolution block*, which does not only increase the spatial resolution but also generates super-resolution (SR) feature maps (S1, S2). These intermediate feature-maps contain high-frequency details, which are lost during the low-resolution (LR) image generation process and used in the *pose estimation block* for better localization. The *super-resolution block* uses a multi-stage design, where each stage increases the spatial resolution of the features maps by a factor of two using the pixel-shuffle algorithm [Shi 2016] (while reducing the number of channels by four). During training, a complete SR image is generated to compute the auxiliary loss **L_HR**, which compares the SR image to the ground truth high-resolution (HR) depth image using the L2 norm. This helps to train the *super-resolution block* and refines the input to the *SR features block*. Note that during training, errors from the pose estimation are also back-propagated to these blocks. Furthermore, at test time only LR images are used and no SR images need to be generated by the network since only the SR feature maps are used.

RTPose was originally developed for color images. Since depth images contain fewer texture details, we have made the architecture more computationally efficient by reducing the number of iterative refinement stages from five to three. The network uses two separate branches, one for keypoint localization and another to compute part affinity maps [Cao 2017]. In our architecture, these two branches consume the 3 types of features (F, S1, S2), where F are the features extracted from the high-resolution feature maps provided by the super-resolution block. The final skeleton is generated from the part affinity and keypoint localization heatmaps using the bipartite graph matching algorithm presented in [Cao 2017]. Losses in the pose estimation network are used as in [Cao 2017], but now take the input from the SR feature maps (S1, S2). At each stage t, two L2 losses L_B^t and L_C^t are computed from the predicted part affinity/keypoint localization heatmaps (B^t/C^t) and the ground truth heatmaps (B^*/C^*) derived from pseudo labels. All the L_B^t and L_C^t losses are summed together to form the pose estimation loss **L_P**. Finally, the total loss is the sum of **L_HR** and **L_P**. We have chosen to weigh both terms equally as we observe that their magnitudes are similar. The complete network is trained end-to-end jointly for both super-resolution and pose estimation.

4.2.2.2 Top-down: ORPose-Depth(krcnn)

We choose the Keypoint R-CNN [He 2017] model as a top-down model illustrated in figure 4.2 (c). We refer to this model as ORPose-Depth(krcnn) tailored to joint person detection and pose estimation. It works as follows: it first extracts the image features using a feature pyramid network (FPN) [Lin 2017a] with a Resnet-50 backbone [He 2016]. The extracted features pass through a region proposal network (RPN) to generate the

bounding-box proposals. The *RoiAlign* layer [He 2017] uses these proposals to extract the fixed-size feature maps. The fixed size feature maps pass through three heads: bounding box head and keypoint head. The bounding box head classifies and regresses for the person bounding box, and the keypoint head generates the spatial heat-maps corresponding to each body keypoint. We use the same multi-task losses as described in [He 2017]. Overall, the loss term L consists of five losses: binary cross-entropy loss for RPN proposal classification $\mathbb{L}_{cls}^{rpn}$, L1 loss for RPN proposal regression $\mathbb{L}_{reg}^{rpn}$, cross-entropy loss [Ross 2017] for bounding box classification $\mathbb{L}_{cls}^{bbox}$, smooth L1 loss for bounding box regression $\mathbb{L}_{reg}^{bbox}$, and cross-entropy loss for the keypoint head $\mathbb{L}_{ce}^{kps}$.

$$L = \sum_i \mathbb{L}_{cls}^{rpn}(x_i^l, y_i^l) + \mathbb{L}_{reg}^{rpn}(x_i^l, y_i^l) + \mathbb{L}_{cls}^{bbox}(x_i^l, y_i^l) + \mathbb{L}_{reg}^{bbox}(x_i^l, y_i^l) + \mathbb{L}_{ce}^{kp}(x_i^l, y_i^l) \quad (4.1)$$

To train the ORPose-Depth(krcnn), we first convert the single-channel depth image to the three channels RGB image by applying the *OCEAN* colormap from OpenCV library[1]. We then propose to exploit advanced data augmentations to adpat the model to the low-resolution depth images. Specifically, we choose a random downscaling factor between 1x to 12x. Then we apply down-sampling and up-sampling operation on the input depth image. The down-sampling operation generates the privacy-preserving low-resolution image, and the up-sampling operation gives the appropriate input resolution needed to train the model. We also use rand-augment [Cubuk 2020], random cut-out [DeVries 2017], and random-flip for the effective regularization.

4.3 Experiments and Results

4.3.1 Training setup

We use the *MVOR-unlabeled* dataset of 80k images and the pseudo labels truth described in Section 4.2.1 for training. When downsampling the images to sizes 80x60 (8x) and 64x48 (10x), we use bicubic interpolation. To generate pseudo labels truth, we use a threshold of 0.7 in the person-detector stage and then select the skeleton if at least 4 keypoints are detected with a score greater than 0.35. We use PyTorch deep learning framework in our experiments. The depth images are normalized in the range [0, 255]. We train the bottom-up models using the stochastic gradient descent optimizer with a momentum of 0.9. The initial learning rate is set to 0.001 with a step decay of 0.1 after 12k iterations and each model is trained for 32k iterations with a batch size of 12. We use the pre-trained weights from the authors of RTPose to initialize the pose-estimator networks. Note that these weights were originally obtained using the color images from the COCO dataset. For the layers that have been modified in the pose-estimation network and contain a larger number of channels (e.g. to accommodate S1 and S2), we repeated the same weights and perturbed them by a small random number. The weights of the super-resolution network are initialized using orthogonal initialization [Saxe 2013]. The top-dwon model, ORPose-Depth(krcnn), is trained on four V100 GPUs with batch

[1] https://docs.opencv.org/4.5.0/d3/d50/group__imgproc__colormap.html

size of 16 (4 images/GPU) and learning rate of 0.001 for 65k iterations using detectron2 framework[2].

Table 4.1: Results of our proposed method (ORPose-Depth(RT) and ORPose-Depth(krcnn)) compared to the baselines(RTPose and *source-only*) for different image resolutions on the *MVOR+* dataset. The *source-only* results correspond to the model evaluated on the color images of the *MVOR+*.

	2D Pose Estimation		**3D Pose Estimation**	
	$\mathbf{AP^{bb}_{person}}$	$\mathbf{AP^{kp}_{person}}$	**MPJPE**	
3D lifter →			*depth*	*2Dto3D*
Models ↓				
source-only				
RTPose	-	33.38	410.00	519.02
Keypoint-RCNN	52.02	45.67	264.61	144.15
Mask-RCNN+Simple-BL	48.74	54.59	261.67	135.81
RTPose_1x	18.02	30.19	392.24	411.76
RTPose_8x	16.79	26.49	405.90	431.29
RTPose_10x	14.48	21.56	416.43	444.56
ORPose-Depth(RT)_8x	18.71	31.68	395.48	414.50
ORPose-Depth(RT)_10x	19.51	30.78	394.40	401.91
ORPose-Depth(krcnn)				
1x	48.94	50.98	256.04	139.68
8x	47.64	47.21	254.87	140.77
10x	46.88	45.84	255.40	141.90
12x	45.82	43.63	258.10	142.88

4.3.2 Testing setup

We evaluate our method on the *MVOR+* dataset. During testing of bottom-up models, we use the flip-test, namely average the original heatmaps with the heatmaps obtained after flipping the images horizontally to refine the predictions. We use the AP^{kp}_{person} for the evaluation of 2D HPE and MPJPE error for the evaluation of 3D HPE as explained in 3.3.3.

4.3.3 Results

We show our results in Table 4.1. The *source-only* results correspond to the evaluation of the default models on the color images of *MVOR+*. As these methods are originally designed for the color images, the aim is to observe how better these models adapt to

[2] https://github.com/facebookresearch/detectron2

the depth images than the counter-part color images. The RTPose and Keypoint-RCNN are the models we use in our training, and Mask-RCNN+Simple-BL is the two-stage model we use to obtain the pseudo labels.

RTPose_1x, RTPose_8x, and RTPose_10x are baseline RTPose models that do not use any super-resolution and are trained on 1x (full-size), 80x60, and 64x48 size depth images, respectively. These RTPose variants are the original models modified to take a 1-channel input. The low-resolution 80x60 and 64x48 images are resampled to the original size using bicubic interpolation to match the input size of the network. The ORPose-Depth(RT)_8x and ORPose-Depth(RT)˙10x are our proposed bottom-up approaches directly trained on 80x60 and 64x48 low-resolution images.

The ORPose-Depth(krcnn) is our proposed top-down approach trained on low-resolution images with a downsampling factor from 1x to 12x. We evaluate the ORPose-Depth(krcnn) on 1x, 8x, 10x, 12x downsampled images. Images are upsampled to the original size before feeding to the ORPose-Depth(krcnn) model.

4.3.3.1 Person bounding box detection

As the top-down approaches, RTPose and ORPose-Depth(RT), do not directly regress for the person bounding box, we evaluate these approaches by fitting a tight bounding box around the keypoints. The top-down approaches regress for the person bounding boxes by design, and as shown in the table 4.1 the top-down approaches perform significantly better than the bottom-up approaches.

Results show that the ORPose-Depth(RT)_8x and ORPose-Depth(RT)_10x models improve by over 1.9% and 5% compared to the baseline RTPose_8x, RTPose_10x models, respectively. We observe a decrease of 3.0% when evaluated at 12x resolution compared to 1x resolution. Compared to *source-only* evaluation, we observe a similar performance for ORPose-Depth(krcnn) at 1x and a decrease in accuracy of around 3% at 12x compared to the pseudo label generator model (Mask-RCNN+Simple-BL). We however observe a decrease of around 3.0% at 1x and 6.0% at 12x for ORPose-Depth(krcnn) compared to Keypoint-RCNN *source-only* baseline. This is likely due to the better bounding box results of Keypoint-RCNN compared to Mask-RCNN as Keypoint-RCNN is trained specifically for a single person class, and Mask-RCNN is trained for 80 COCO classes.

4.3.3.2 2D human pose estimation

Results show that the ORPose-Depth(RT)_8x and ORPose-Depth(RT)_10x models improve by over 5% and 9% compared to the baseline RTPose_8x, RTPose_10x models, respectively. More interestingly, Both ORPose-Depth(RT)_8x, ORPose-Depth(RT)_10x perform better than RTPose_1x model. We attribute these improvements to our proposed design improvements in the architecture compared to the full-size RTPose_1x model. The ORPose-Depth(krcnn) model performs significantly better than the ORPose-Depth(RT) models on all resolutions and improves the *source-only*

Keypoint-RCNN model over 5% at 1x, 1.5% at 8x, and 0.1% at 10x. The AP result for the pseudo label generator model (Mask-RCNN+Simple-BL) on the color images is 54.59, showing that there still exists a gap of around 4% at 1x and around 11% at 12x to be filled between the depth and color images.

4.3.3.3 3D human pose estimation

As shown in 4.1, lifting the 2D keypoints to 3D using an *off-the-shelf 2Dto3D* lifting network gives better results compared to lifting them from the depth images. The bottom-up approaches perform much more poorly compared to top-down approaches, likely due to poor keypoint localization, which is used in the MPJPE calculation to match the ground truth person with the detected person using glsoks metric. Figure 4.3 shows qualitative results comparing our proposed approach with the baselines. Additional qualitative results are available in the illustrative video[3]

4.4 Conclusion

In this chapter, we present an approach for high-resolution multi-person 2D pose estimation from low-resolution depth images. Our evaluation on the *MVOR+* dataset shows that even with a 12x subsampling of the depth images, our method achieves results equivalent to a pose estimator trained and tested on the original-size images. These results suggest the high potential of low-resolution images for scaling up and deploying privacy-preserving AI assistance in hospital environments. Furthermore, we show that by exploiting high-quality pose detections on the color images of a non-annotated RGB-D dataset, we can generate pseudo labels for the depth images and train a decent OR pose estimator. We further show that using improved underlying HPE architecture and strong data augmentations significantly boosts the performance and effectively learns the features for low-resolution images. These experimental observations lay the foundations for designing the unsupervised domain adaption approach on the color images for our next chapter.

[3] https://cutt.ly/depthpose

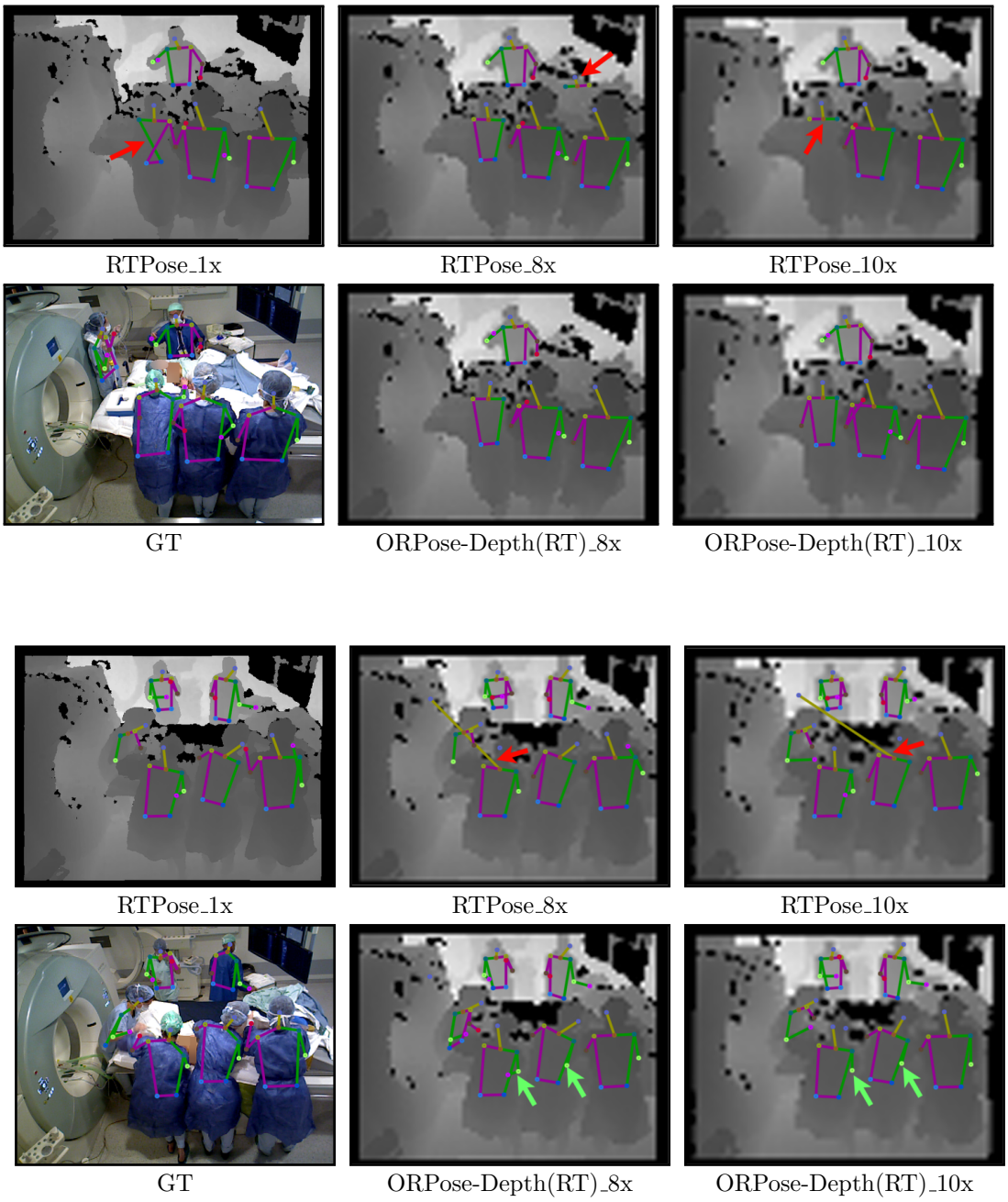
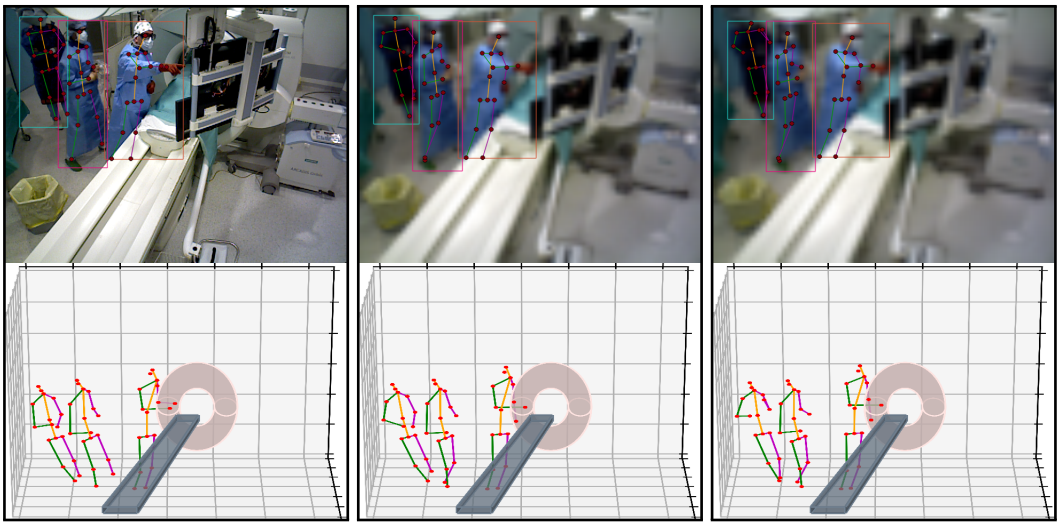


Figure 4.3: ORPose-Depth(RT)_8x and ORPose-Depth(RT)_10x, w.r.t the baseline models RTPose_1x, RTPose_8x, and RTPose_10x. We also show the labels truth (GT) on color images for better appreciation of the qualitative results. These results show that ORPose-Depth(RT)_8x and ORPose-Depth(RT)_10x perform better for removing false positives and spurious detections and improve the part localization (see red and green arrows in the figures)

5 Self-supervision on unlabelled OR color images for joint 2D/3D human pose estimation

Self-supervised learning is the cake, supervised learning is the icing on the cake, reinforcement learning is the cherry on the cake.
– Yann LeCun



1x 8x 10x

A sample qualitative result from our unsupervised domain adaptation approach on different low-resolution color images for multi-person 2D/3D human pose estimation. Demo video is available here: https://cutt.ly/orpose3d. Project page: https://github.com/CAMMA-public/ORPose-Color

Chapter Summary

5.1 Introduction

Following the motivation from the last chapter that devises an unsupervised domain adaptation approach for adapting a model trained on the color images to the depth images, in this chapter, we propose an unsupervised domain adaptation approach on Operating Room (OR) color images for real-time 2D/3D Human Pose Estimation (HPE). As outlined in the section 3.3, the current deep learning approaches for HPE employ multi-stage deep neural networks to achieve state-of-the-art performance. For example, top approaches for the task of 2D pose estimation on the COCO dataset use a two-stage approach, in which the first stage determines the person's bounding boxes and the second stage estimates the keypoints for each bounding box. This multi-stage design using high-capacity deep neural networks in both stages helps to achieve better accuracy. The run-time performance of such a design is however considerably low. More practical solutions such as Mask-RCNN [He 2017] or RTPose [Cao 2017] use low-capacity and single-stage networks for the same task to achieve better run-time performance, but the accuracy of these system is low compared to the multi-stage systems. Therefore, one key challenge for the deployment of HPE network inside the OR is not only to give accurate predictions, but also to be fast using a light-weight and single-stage end-to-end design, as needed for real-time applications.

In this chapter, we work at the intersection of knowledge distillation [Hinton 2015, Zhang 2019a] and data distillation [Radosavovic 2018] and exploits these techniques to solve the task of multi-person 2D/3D HPE without using any manual annotations. We use knowledge distillation to transfer knowledge from an accurate, larger, and multi-stage teacher network to a practical, smaller, and single-stage student network. The idea of knowledge distillation has been adapted to different problems. In [Hinton 2015], authors use the probability output vector of a teacher

network as soft-labels to train a student network for multi-class classification. The student network learns jointly from the soft-labels generated by the teacher output and from the hard-labels given by the ground truth. Similarly, in [Zhang 2019a], authors use this approach for single-person 2D HPE by jointly training the student network from the soft output heatmaps of a teacher network and the hard ground truth heatmaps. Both the student and the teacher network work on the fully supervised dataset, and the soft output of the teacher network serves as additional useful labels along with the hard labels obtained from the supervised ground truth.

We aim at applying knowledge distillation when only *non-annotated* data is available in the target domain. Instead of using supervised annotations, we propose to use data distillation to generate labels automatically from the non-annotated dataset. We run a complex teacher network which ensembles output predictions on geometrically transformed input images. Unlike the standard use of data distillation [Radosavovic 2018], which only exploits hard predictions obtained by removing the low confidence keypoints, we also use the soft predictions from the confidence value for each keypoint.

As student network, we use a low capacity single stage network based on Keypoint-RCNN. The architecture of the student network is inspired from [Dabral 2019] and further extended to effectively use the *hard-set* and *soft-set* for joint 2D/3D multi-person HPE in the OR. By utilizing our approach, the student network reaches an accuracy on par with the teacher network.

As discussed in section 1.2.1, another specific issue in the OR is to preserve the privacy of patients and clinicians while performing computer vision tasks. Human pose estimation on low-resolution images has been suggested to improve the privacy as discussed in the previous chapter. We therefore also extend our approach to deliver accurate poses on low-resolution images with downsampling factor as low as 12x.

5.2 Methodology

5.2.1 Problem overview

Given a monocular RGB input image $\mathcal{I}$ of size $W \times H$, our task is to detect the 2D and 3D body keypoints for multiple persons using a single efficient end-to-end network. The 2D keypoints $\mathcal{P}_{2D} \in \mathbb{R}^{m \times n \times 2}$ are in image coordinates, and the 3D keypoints $\mathcal{P}_{3D} \in \mathbb{R}^{m \times n \times 3}$ are in the root-relative coordinates (where the root-joint is set to be the origin and all other joints are measured w.r.t root-joint). Here, m is the number of persons, and $n = 17$ is the number of joints for each pose. We also consider low-resolution images for the same task, which have very small input sizes. To tackle the problem, we utilize a teacher/student approach to train the end-to-end student network by distilling the knowledge in a teacher network on large-scale unlabelled data. We follow a two-step approach: in the first step, a multi-stage high-capacity neural network (a teacher network) is used to generate pseudo labels; in the second step, these pseudo labels are used to train an end-to-end low-capacity network (a student network).

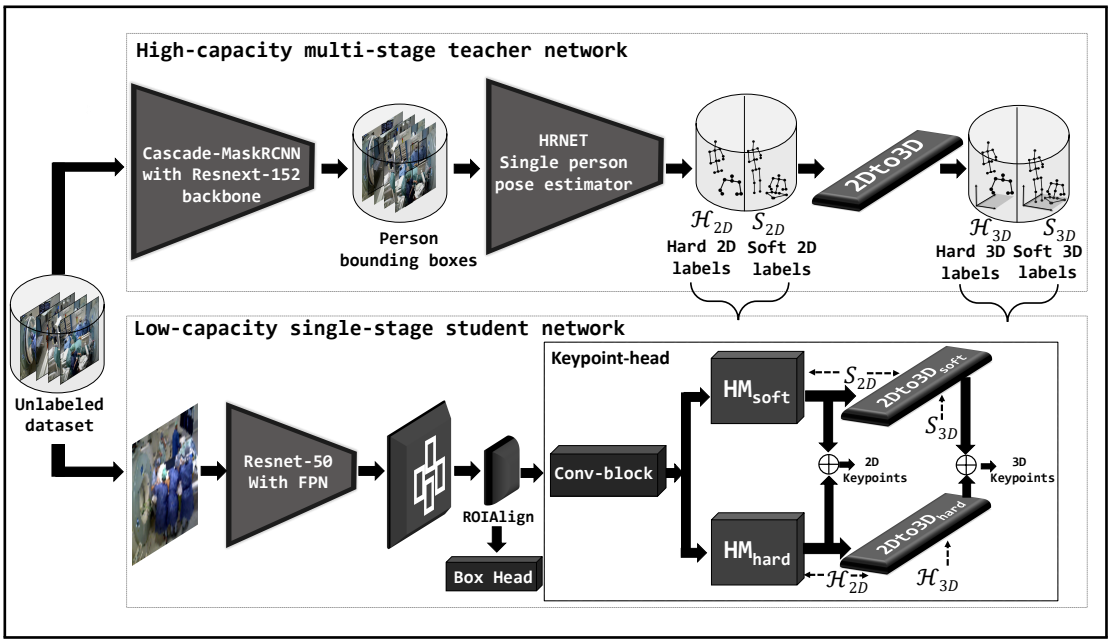


Figure 5.1: Proposed self-supervised methodology for joint 2D/3D keypoint estimation using the teacher/student paradigm. The teacher network is a three-stage network which uses the unlabelled dataset to extract person bounding boxes, estimate 2D keypoints, and regress 2D keypoints to 3D. It generates soft and hard pseudo labels to be used by the student network. The student network is a single-stage network and effectively utilizes the soft and hard pseudo labels to jointly estimate the 2D and 3D keypoints.

5.2.2 Knowledge generation using the teacher network

The teacher network, shown in figure 5.1, is a three-stage network: The first stage uses the cascade-mask-rcnn [Cai 2019b] with the resnext-152 [Xie 2017] backbone to generate person bounding boxes, the second stage estimates the 2D keypoints for each bounding box using the HRNet architecture [Sun 2019] after discarding low-score bounding boxes, and the third stage lifts the detected 2D keypoints to the 3D using a residual-based 2-layer fully-connected network [Martinez 2017]. The three stages in the teacher network are selected based on their state-of-art performance on the COCO and Human3.6 dataset. The first and second stages are trained on the COCO dataset [Lin 2014] and the third stage is trained on the Human3.6 dataset [Ionescu 2013]. Multi-level scaling and flipping transformations are applied in the first and the second stage to obtain good quality person bounding boxes and 2D keypoints. However, errors can still be present in the keypoints and are encoded in the keypoint confidence scores. Therefore, we propose to construct two sets of pseudo-labels: the soft-set $\mathcal{S}$ and the hard-set $\mathcal{H}$. The soft-set $\mathcal{S} = \{\mathcal{S}_{2D}, \mathcal{S}_{3D}\}$ consists of soft 2D keypoints and soft 3D keypoints. Soft 2D keypoints $\mathcal{S}_{2D} \in \mathbb{R}^{m \times n \times 3}$ are obtained by storing the confidence value for each keypoint along with their coordinates. The last dimension in $\mathbb{R}^{m \times n \times 3}$ represents the channel for the confidence value. $\mathcal{S}_{2D}$ is sent to the third stage to obtain the soft 3D keypoints $\mathcal{S}_{3D}$. Similarly, the hard-set $\mathcal{H} = \{\mathcal{H}_{2D}, \mathcal{H}_{3D}\}$ consists of hard 2D keypoints and hard 3D

keypoints. $\mathcal{H}_{2D}$ is obtained by only keeping the high confidence 2D keypoints and discarding the low confidence keypoints. $\mathcal{H}_{3D}$ is obtained by passing $\mathcal{H}_{2D}$ to the lifting network. We show in the experiments that these two sets provide useful learning signals when used to train the student. In the next section, we show how we exploit these two sets of pseudo labels for effectively training the student network.

5.2.3 Knowledge distillation in the student network:

The student network presented in figure 5.1 is an end-to-end network based on Keypoint-RCNN that jointly predicts the 2D and 3D poses. We replace the mask head of the Mask-RCNN network with a keypoint-head for joint 2D and 3D pose estimation. The keypoint-head accepts the fixed size proposals from the ROIAlign layer and passes them through 8 conv-block layers to generate the features. These features are upsampled using a deconv and bi-linear upsampling layer into two branches to generate 17 channel heatmaps corresponding to each body joint. The first branch upsamples the features to generate the heatmaps HM_{soft}, and the second branch upsamples them to generate the heatmaps HM_{hard}. The HM_{soft} and HM_{hard} heatmaps are connected to their respective lifting networks i.e $2Dto3D_{soft}$ and $2Dto3D_{hard}$ to lift the incoming 2D keypoints to 3D.

Training: Training of the network follows the same framework as Mask-RCNN along with the additional losses coming from the keypoint-head. In the keypoint-head, we compute 2D and 3D losses L_{2D} and L_{3D} to estimate the 2D and 3D keypoints. L_{2D} consists of soft and hard 2D keypoint losses. The soft 2D keypoint loss L_{2Dsoft} is obtained by first multiplying HM_{soft} with the corresponding confidence values from the last channel of $\mathcal{S}_{2D}$ and then computing its cross-entropy loss with $\mathcal{S}_{2D}$. The hard 2D keypoint loss L_{2Dhard} is obtained by calculating the cross-entropy loss between HM_{hard} and $\mathcal{H}_{2D}$. Similarly, the 3D loss L_{3D} consists of soft and hard 3D keypoint losses. Soft 3D keypoint loss L_{3Dsoft} is obtained by taking the smooth L1 loss between $\mathcal{S}_{3D}$ and the output of $2Dto3D_{soft}$ using the input $\mathcal{S}_{2D}$, and hard 3D keypoint loss L_{3Dhard} is obtained by taking the smooth L1 loss between $\mathcal{S}_{3D}$ and the output of $2Dto3D_{hard}$ using the input $\mathcal{H}_{2D}$. All four losses are added together to obtain the loss for the keypoint-head L_{kpt}. The overall loss is the sum of L_{kpt} with the standard Faster-RCNN loss, ie. the bounding box classification and regression loss, and the region proposal loss.

Inference: During inference, the 2D keypoints are computed by taking the arg-max over each channel from the mean output of HM_{soft} and HM_{hard}, and the 3D keypoints are computed by calculating the 2D keypoints from HM_{soft} and HM_{hard} using arg-max, passing the 2D keypoints to the respective *2Dto3D* lifting network, and averaging the 3D output.

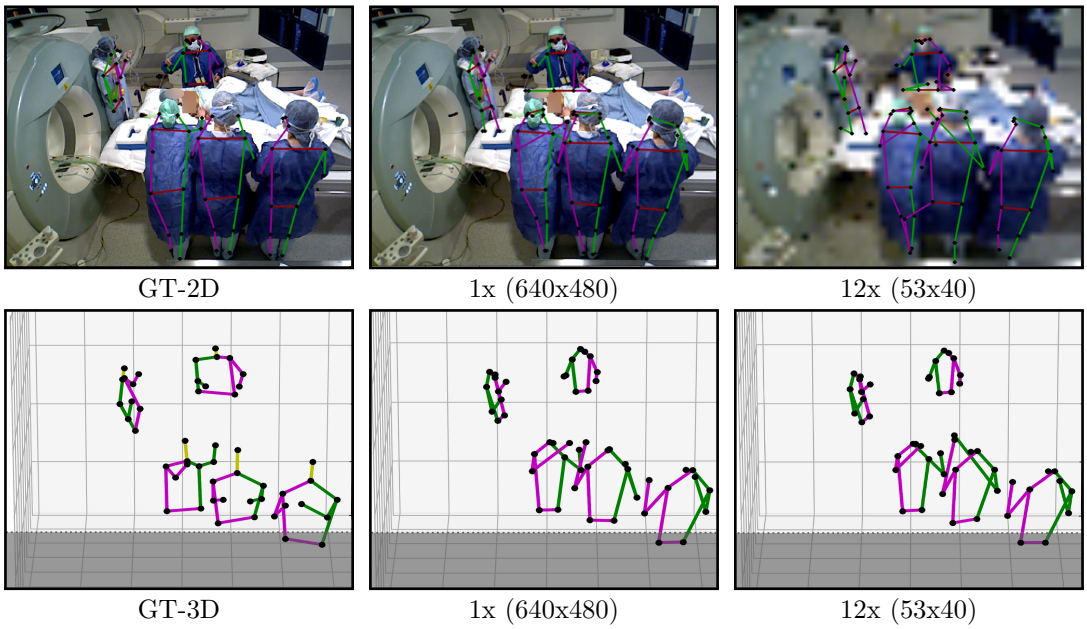


Figure 5.2: Qualitative results for 2D and 3D keypoints estimation from the student network (ORPose_all) at original and downsampled image sizes. GT-2D and GT-3D are the visualization results from 2D and 3D ground truth keypoints respectively. Since we are not predicting the scale of the 3D pose in the camera frame, we use the depth of the root node from the ground truth as scale to generate this visualization.

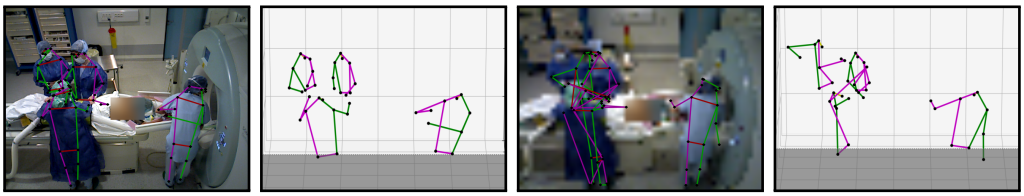
5.3 Experiments and results

5.3.1 Training and testing dataset

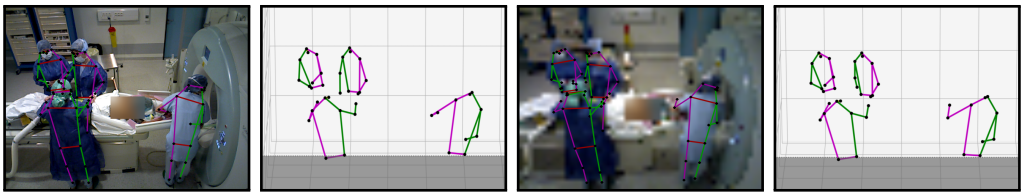
We use *MVOR-unlabeled* dataset to generate pseudo labels and train our networks and the *MVOR+* dataset as a test set as described in section 3.2.1 and 3.2.2.2, respectively. Since we are evaluating 3D poses from the *MVOR+* dataset for the single view, we projected these 1061 3D poses into the respective camera coordinates to obtain 2926 valid 3D poses (we discarded the not-visible poses). The original size of all the images is 640x480. We also conduct experiments with downsampled images using the scaling factors 8x, 10x, and 12x, yielding images of size 80x64, 64x48, and 53x40. We use AP^{bb}_{person} for person bounding box evaluation, AP^{bb}_{person} for 2D keypoint evaluation, and MPJPE error for 3D keypoint evaluation, respectively, as described in section 3.3.3.

5.3.2 Experiments

The student network is trained differently for two sets of experiments, yielding the networks ORPose_fixed_s**x** (**s**=1,8,10,12) and ORPose_all. ORPose_fixed_**s**x is trained using either images of the original size (**s**=1) or low-resolution images at a fixed scaling factor (**s**=8,10,12). When feeding the networks, low-resolution images are first upsampled to match the original input size.
Evaluation of networks ORPose_fixed_**s**x is done on the same scale they are trained on.



Results of default student network for 1x and 12x downsampling scale



Results of trained student network for 1x and 12x downsampling scale

Figure 5.3: Comparative qualitative results for the default and the trained student networks. (a) The default student network uses the pre-trained COCO and Human3.6 weights. (b) The trained student network exploits the soft and hard pseudo labels obtained from the teacher network. The left side shows the 2D/3D visualization results at 1x scale, and the right side shows the 2D/3D visualization results at 12x scale.

In the second experiment, ORPose_all is trained using original and downsampled images with a random downsampling factor. This is similar to the scaling data augmentation technique, but we consider here a very low-resolution scenario where the input image is downsampled up to 12x. We choose the random scale such that for 30% of the training time there is no downsampling, for 35% the downsampling scale is randomly chosen between 2 and 8, and for the remaining 35% of training time downsampling scale is randomly chosen between 8 and 12. The intention to train the network using this strategy is to obtain a single model that can work on high-resolution images and should also perform considerably better on the low-resolution images. The base learning rate for ORPose_fixed_1x is set to 1e-3 for 5k total number of iterations with a step decay of 0.1 after 2k, 3k, and 4k iterations; the base learning rate for ORPose_fixed_**s**x (**s**=8,10,12) is set to 1e-2 for 10k total number of iterations with a step decay of 0.1 after 7k, 8k, and 9k iterations; the base learning rate for ORPose_all is set to 1e-1 for 20k total number of iterations with step decay of 0.1 after 14k, 16k, and 18k iterations. The downsampling and upsampling operation is performed using bilinear interpolation. We use a detectron2 framework [Wu 2019a] to run all the experiments on two V100 NVidia GPUs using the distributed data parallelism framework of PyTorch. We use a batch size of 32 and the stochastic gradient solver as the optimizer for all the experiments.

5.3.3 Results

Table 5.1 shows the results for the teacher and the student networks on *MVOR+*, along with the network parameter complexity, before training on OR data. These networks are initialized from the COCO and Human3.6 pre-trained network weights. We evaluated both on the original image size (1x) and downsampled images of scale 8x, 10x, and 12x. As shown in the Table 5.1, there exists a margin of 11.6% 2D keypoint AP. Also, the 3D

Network	#Params	GFLOPs	Scale	AP^{bb}_{person}	AP^{kp}_{person}	*MPJPE*
Teacher	250.1M	1048.8	**1x**	**53.81**	**57.78**	**134.88**
			8x	39.03	29.28	170.25
			10x	31.90	18.60	203.89
			12x	24.38	8.89	260.83
Student	67.9M	215.0	**1x**	**52.77**	**46.17**	**147.17**
			8x	40.21	27.02	168.10
			10x	34.12	20.06	181.69
			12x	29.19	14.42	194.35

Table 5.1: Baseline results on *MVOR+* for teacher and student networks when no training is performed on OR data. Higher AP and lower MPJPE are better. Student and teacher networks are evaluated at original and low-resolution sizes. The aim is to train the student to reach the same performance as the teacher at high resolution (1x).

Student Network	Scale	AP^{bb}_{person}	AP^{kp}_{person}	*MPJPE*
ORPose_fixed_1x	1x	**50.87**	55.20	134.23
ORPose_fixed_8x	8x	49.50	53.50	137.40
ORPose_fixed_10x	10x	49.01	51.98	137.71
ORPose_fixed_12x	12x	48.23	49.88	138.83
ORPose_all	1x	50.59	**55.80**	**134.13**
	8x	49.57	53.31	136.45
	10x	49.25	52.12	136.95
	12x	47.54	49.51	138.35

Table 5.2: Results of our student network evaluated at original size and low resolution images. ORPose_fixed_**s**x (**s**=1,8,10,12) are trained and evaluated at fixed scale. ORPose_all is a single model trained on random size low resolution and high resolution images, and evaluated on original size images and fixed scale downsampled images.

Student Network	Scale	AP^{bb}_{person}	AP^{kp}_{person}	*MPJPE*
Single-branch(hard)	1x	50.61	54.73	145.77
Single-branch(soft)	1x	**51.04**	54.70	134.20
Single-branch(hard+soft)	1x	50.95	55.11	152.28
Double-branch(hard+soft)	1x	50.59	**55.80**	**134.13**

Table 5.3: Ablation study on the student network, by comparing to a single branch trained using hard, soft and hard+soft labels. We achieve the best result when using our proposed two-branch design for both 2D and 3D keypoint estimation.

error in the student network is 12.30 mm more compared to the teacher network. When we evaluate these models on the low-resolution images, we observe a strong decrease in the performance, likely because such low-resolution images were not much represented in the training dataset. The low-resolution results of the teacher network are somewhat worse compared to the student network, possibly due to the multi-stage design of the teacher network, where the poor performance of the current stage affects the next stage. The student network is less affected, likely due to its single-stage design. We believe the

this due to the fact that the distribution of input images on these extreme low-resolutions changes considerably which results in such a large decrease in the performance.

Table 5.2 shows the results for our student network when trained using the soft and hard pseudo labels for 2D/3D keypoints obtained from the teacher network. We observe improved performance in all the models when trained with the pseudo labels. ORPose_all achieves nearly the same performance compared to the models trained for specific scale low-resolution images. Performance of ORPose_all on the high-resolution images nearly reaches the teacher network and on the low-resolution images this network performs much better. This is illustrated in the qualitative results shown in figure. 5.2 and figure. 5.3. This suggest single network with better training paradigm is able to learn the pose details in domain specific scenario for example the operating room in our case.

5.3.4 Ablation study

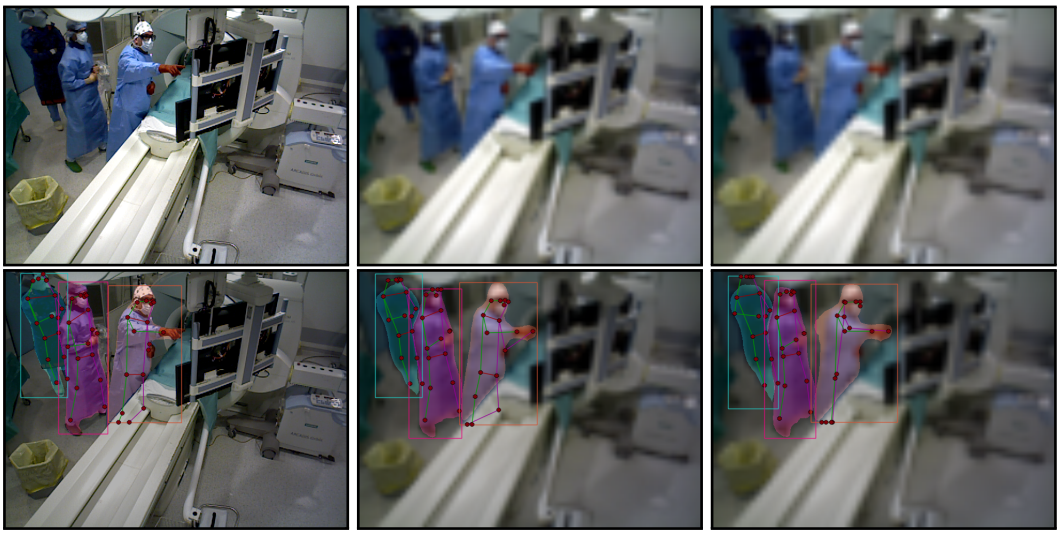
To evaluate the effect of soft-labels on the student network, we keep only one branch for 2D/3D keypoint estimation i.e only one heatmap layer and one *2Dto3D* network. We train this single branch keypoint-head with only hard labels, only soft labels, and both hard and soft labels. To train for both the hard and soft labels, the 2D losses are computed using the same heatmap layer and 3D losses are computed using the same *2Dto3D* network. As shown in Table 5.3, we observe that training with the hard labels hurts the 3D keypoint estimation, and training using only the soft labels achieves good overall results. 2D keypoint estimation is however inferior compared to our two-branch design trained for soft and hard losses.

5.4 Conclusion

In this chapter, we tackle joint 2D/3D pose estimation from monocular RGB images and propose a self-supervised approach to train an end-to-end and easily deployable model for the OR. We use data distillation to exploit non-annotated data and knowledge distillation to benefit from the high-quality predictions of a multi-stage high capacity pose estimation model. Our approach does not require any ground truth poses from the OR and evaluation on the *MVOR+* dataset suggests its effectiveness. We further demonstrate that the proposed network can yield accurate results on low-resolution images, as needed to ensure privacy, even using a downsampling rate of 12x.

6 Unsupervised domain adaptation for person pose estimation and instance segmentation in the OR

Instead of trying to produce a programme to simulate the adult mind, why not rather try to produce one which simulates the child's? If this were then subjected to an appropriate course of education one would obtain the adult mind.
– Alan Turing, 1950, "Computing Machinery and Intelligence"



1x 8x 10x

A sample qualitative result from our unsupervised domain adaptation approach on different low-resolution color images for joint person pose estimation and instance segmentation. Demo video: https://youtu.be/gqwPu9-nfGs, project page: https://github.com/CAMMA-public/HPE-AdaptOR

Chapter Summary

6.1 Introduction

In this chapter, we propose a novel Unsupervised Domain Adaptation (UDA) approach, called *AdaptOR*, for joint person pose estimation and instance segmentation. We aim to adapt a model from a labeled source domain, i.e., unconstrained natural images from *COCO* [Lin 2014] to an unlabeled target domain, i.e., constrained low-resolution OR

images downsampled at different resolutions from 1x to 12x. The UDA methods have been extensively studied for various computer vision tasks ranging from image classification [Zhuang 2020], object detection [Oza 2021] to semantic segmentation [Toldo 2020]. Unlike the existing UDA approaches that have primarily been applied to general object classes, we aim to study the UDA for a single but highly challenging "person" class inside the visually complex OR environment while simultaneously exploiting the articulated "person" class properties for effective domain adaptation.

We choose Mask R-CNN [He 2017] as our backbone model for joint person pose estimation and instance segmentation, which is primarily designed for a single domain fully supervised training. Inspired from UDA for image classfication [Chang 2019], we propose *disentangled feature normalization* (DFN) for our backbone model to train it on two statistically different domains. DFN replaces every feature normalization layer in the feature extractor of the backbone model with two feature normalization layers: one for the source domain and another for the target domain. With the improved design, the backbone model expects an input batch containing half the images from the source domain and another half from the target domain. DFN therefore modifies the multi-task loss function to compute and weigh the loss differently for the two domains. The use of separate feature normalization layers for the two domains effectively disentangle the feature learning and stabilizes the training.

Given a backbone model with the ability to train on two statistically distinct domains, we build our approach based on a *self-training* framework [Sohn 2020a, Liu 2021b, Deng 2021], where we aim to predict similar predictions from a model under different augmentations of the same image, thereby taking the confident predictions from one augmented image - called *weakly* augmented image - as pseudo labels for the other augmented image - called *strongly* augmented image. The idea has primarily been utilized for the image classification tasks [Berthelot 2019a, Sohn 2020a] where the model predictions need to be invariant to the different augmentations applied to the input image. The spatial localization tasks such as pose estimation or instance segmentation, however, can change the model predictions under certain geometric augmentations, e.g., random-flip or random-resize. Thankfully, these changes in the predictions need to satisfy *transformation equivariant constraints*, i.e. prediction labels also need to be transformed according to the applied geometric augmentations. We therefore use the *transformation equivariant constraints* to add explicit geometric constraints on the *weakly* and the *strongly* augmented unlabeled images to generate high-quality pseudo labels; for example, the random-flip operation has to exploit the chirality property [Yeh 2019] for pose estimation to map the keypoints to the horizontally flipped image.

To improve the performance of the model on low-resolution OR images as needed for privacy preservation, we also propose to extend the data augmentation pipeline with a *strong-resize* augmentation for the *strongly* augmented image by applying two resize operations on the input image: a down-sampling and an up-sampling operation with a

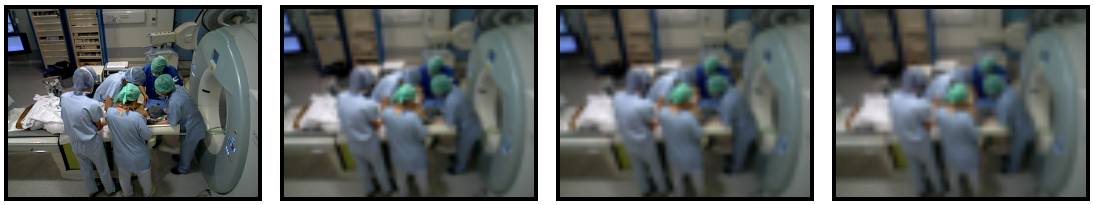
scaling factor randomly chosen between 1x to 12x. It generates heavily blurred images (see the example downsampled images in figure 6.1) that naturally extend our approach to the privacy-preserving low-resolution images. Training the model using the two sets of weak and strong augmentations also enforces consistency regularization [Tarvainen 2017, Sajjadi 2016, Sohn 2020a]: a popular regularization technique utilized in a Semi-supervised Learning (SSL). The SSL is closely related to the UDA and aims to generalize a model to the same domain with limited labeled and large-scale unlabeled data.

We further extend our approach with *mean-teacher* for stable training [Tarvainen 2017], where instead of using a single model to generate and consume the pseudo labels, we create two copies of a given source domain trained model: a *teacher* and a *student* model. The *teacher* model generates the pseudo labels on the *weakly* augmented image that is used by the *student* model to train on the corresponding *strongly* augmented image. The weights of the *teacher* model are updated using temporal ensembling of the weights of the *student* model, thereby helping it to improve its predictions due to ensembling while simultaneously generating better pseudo labels to improve the *student* model. Figure 6.2 illustrates the complete architecture of our approach.

We evaluate our approach on the two OR datasets: *MVOR+* and *TUM-OR-test* [Belagiannis 2016]. The default annotation in the *TUM-OR-test* contains only the six common COCO keypoints in the upper body bounding box. Therefore, we re-annotate the *TUM-OR-test* using a semi-automatic approach[1]. Both *MVOR+* and *TUM-OR-test* do not contain ground-truth for the person instance masks. We therefore evaluate the mask segmentation results by computing tight bounding boxes around the prediction masks and comparing them with the corresponding ground-truth bounding boxes, along with qualitative results. We show that our approach performs significantly better after domain adaption and against strongly constructed baselines, especially on privacy-preserving low-resolution OR images even downsampled up to **12x**. As our backbone model based on Mask R-CNN performs person bounding box detection by design, we use the model to evaluate for the person bounding boxes and show significant improvements in the bounding box detection results. We also conduct extensive ablation studies to shed light on the different components of our approach and their contributions to the results. The figure **??** shows a comparative qualitative result before and after the domain adaptation.

Finally, without bells and whistles, our UDA approach can be easily used as an SSL approach on the same domain dataset - by using regular feature normalization instead of DFN. We show the generality of our approach as an SSL method on the same domain COCO dataset with different percentages of supervision. With as few as **1%** of labeled supervision, we obtain 57.7% (38.2 keypoint Average Precision (AP)) in the pose estimation and 72.3% (36.1 mask AP) in instance segmentation, a strong improvement against the model trained with 100% of labeled supervision (66.2 keypoint AP and 49.9

[1]The updated *MVOR+* dataset and the new *TUM-OR-test* annotations along with the source code and models will be available at https://github.com/CAMMA-public/HPE-AdaptOR.



(a) 640x480 (1x) (b) 80x60 (8x) (c) 64x48 (10x) (d) 53x40 (12x)

Figure 6.1: Sample image from the OR downsampled at different resolutions. With significant degradation in the spatial details, these are more suitable for activity analysis in privacy-sensitive OR environments.

mask AP). These initial valuable baselines for the joint person pose estimation and instance segmentation could help foster SSL research on large-scale traditional vision datasets.

6.2 Detailed methodology

6.2.1 Problem overview

Given an end-to-end model for joint person pose estimation and instance segmentation trained on the source domain labeled dataset $\mathcal{X} = \{x_i|y_i\}_{i=1}^{N_l}$, we aim to adapt it to the unlabeled target domain dataset $\mathcal{U} = \{u_j\}_{i=1}^{N_u}$. The source domain images are natural *in the wild* images, whereas the target domain images are the high-resolution and low-resolution (downsampled up to 12x) images from the OR. N_l and N_u are the number of labeled and unlabeled images, respectively. The source domain's labeled dataset consists of images x_i with the corresponding ground-truth labels y_i. The ground-truth labels y_i consist of bounding boxes $\mathcal{P}_{bbox} \in \mathbb{R}^{m\times 4}$, keypoints $\mathcal{P}_{kp} \in \mathbb{R}^{m\times n\times 2}$, and masks $\mathcal{P}_{mask} \in \mathbb{R}^{m\times p\times 2}$, where m is the number of persons, n is the number of 2D keypoints for each pose, and p is the number of contour points on the ground-truth binary mask. The unlabeled data from the target domain consists of only the images u_j.

We first explain the backbone models chosen for this work and the proposed UDA method, which we call *AdaptOR*. Briefly, we first extend Mask R-CNN [He 2017] with disentangled feature normalization (DFN) to handle the statistically different datasets from the two domains. Then we develop our approach by designing geometrically constrained data augmentations to generate and use the pseudo labels for adapting the model to the unlabeled target domain consisting of high- and low-resolution images from the OR.

6.2.2 Backbone models

We choose the Mask R-CNN [He 2017] model, where the mask and the keypoint head are designed to use a single person class. We refer to this model as *km-rcnn* tailored to joint person pose estimation and instance segmentation. It can also perform person bounding box detection by design. *km-rcnn* works as follows: it first extracts the image features using a feature pyramid network (FPN) [Lin 2017a] with a Resnet-50 backbone [He 2016].

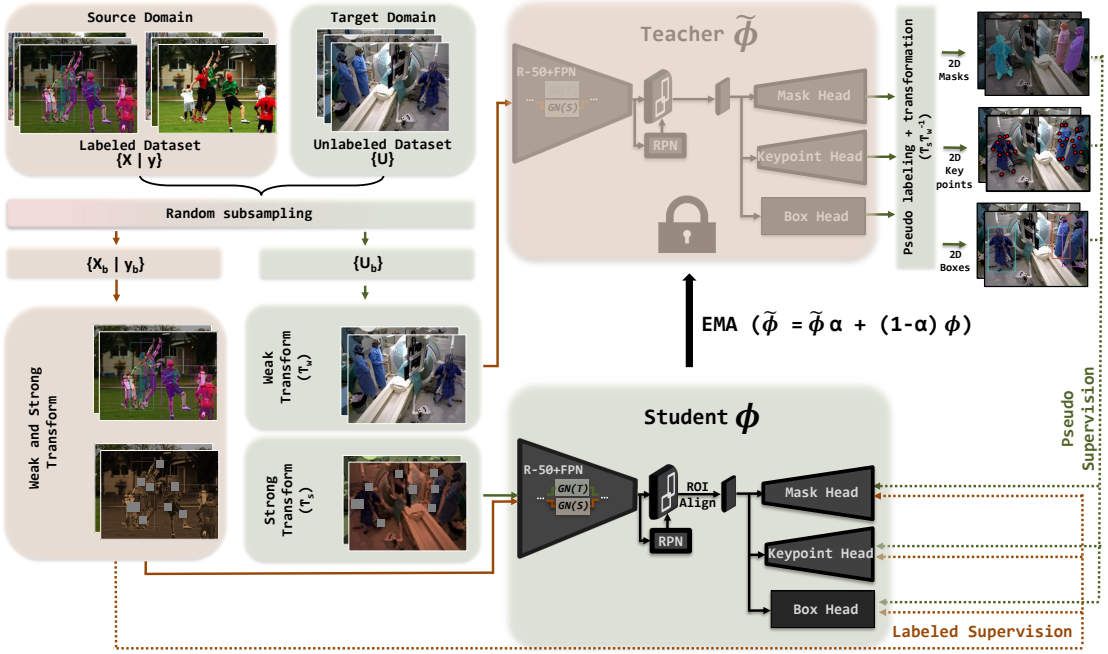


Figure 6.2: Overview of our approach for unsupervised domain adaptation. We generate two types of augmentations on the given unlabeled target domain images: weak and strong. The weakly augmented images pass through a frozen teacher model and a thresholding function to generate the pseudo labels. These pseudo labels are then geometrically transformed to the corresponding strongly augmented image space. A student model uses these transformed pseudo labels to train on the strongly augmented unlabeled images jointly with the labeled source domain images. The weights of the frozen teacher model are updated using the exponential moving average (EMA) of the student model's weights. We also replace every group normalization (GN) layer in the feature extractor with two GN layers (*GN(S)* and *GN(T)*) to normalize features of two domains separately, as needed to handle statistically different source and target domains.

The extracted features pass through a region proposal network (RPN) to generate the bounding-box proposals. The *RoiAlign* layer [He 2017] uses these proposals to extract the fixed-size feature maps. The fixed-size feature maps pass through three heads: bounding box head, keypoint head, and mask head. The bounding box head classifies and regresses for the person bounding box, the keypoint head generates the spatial heat-maps corresponding to each body keypoint, and the mask head generates segmentation masks. We use the same multi-task losses as described in [He 2017] except for bounding box classification loss where we use focal loss [Ross 2017] instead of cross-entropy loss for the better handling of foreground-background class imbalance in our UDA framework [Liu 2021b]. Overall, the supervised loss term $\mathcal{L}_s$ consists of six losses: binary cross-entropy loss for RPN proposal classification $\mathbb{L}_{cls}^{rpn}$, L1 loss for RPN proposal regression $\mathbb{L}_{reg}^{rpn}$, focal loss [Ross 2017] for bounding box classification $\mathbb{L}_{cls}^{bbox}$, smooth L1 loss for bounding box regression $\mathbb{L}_{reg}^{bbox}$, cross-entropy loss for the keypoint head $\mathbb{L}_{ce}^{kps}$, and the binary cross-entropy for the mask head $\mathbb{L}_{bce}$.

$$\mathcal{L}_s = \sum_i \mathbb{L}_{cls}^{rpn}(f_i^l, y_i^l) + \mathbb{L}_{reg}^{rpn}(f_i^l, y_i^l) + \mathbb{L}_{cls}^{bbox}(f_i^l, y_i^l) + \mathbb{L}_{reg}^{bbox}(f_i^l, f_i^l) + \mathbb{L}_{ce}^{kp}(f_i^l, y_i^l) + \mathbb{L}_{bce}^{mask}(f_i^l, y_i^l). \tag{6.1}$$

Here, f_i^l and y_i^l correspond to the features and the ground-truth labels for the labeled input image x_i^l.

6.2.2.1 Initialization

The state-of-the-art approaches for downstream tasks such as object detection [Ren 2015] and instance segmentation [He 2017] initialize the backbone network from the supervised ImageNet [Deng 2009] weights. The feature normalization during the training is performed using frozen batch normalization (BN) in all the feature extraction layers. It, in turn, uses statistics (mean and variance) derived from the ImageNet training set and freezes its affine parameters (weights and biases). The current advancements in self-supervised methods to learn generic feature representations exploiting large-scale unlabeled data have started to surpass the supervised ImageNet baselines on the downstream tasks [Chen 2020, He 2020, Misra 2020]. However, the backbone feature extractor weights from the self-supervised methods may not have the same distribution as supervised ImageNet methods. The use of frozen BN during the training therefore could lead to unstable training. Authors in [He 2020] suggest training the BN layers using Cross-GPU BN [Peng 2018] to circumvent the issue. We find in our experiments that group normalization (GN) [Wu 2018] works equally well without the overhead of communicating the batch statistics over all the GPUs resulting in an increased training speed. We follow the network design from [Wu 2018, Wu 2019c] to change the BN layers of *km-rcnn* with the GN layers. The updated model, called *km-rcnn+*, is initialized from the self-supervised method MoCo-v2 [Chen 2020, He 2020] and trained on the source domain dataset.

6.2.2.2 Disentangled feature normalization

Given the model, *km-rcnn+*, trained on the labeled source domain dataset, we aim to adapt it to the unlabeled target domain. We observe in our experiments that feature normalization plays a vital role in training the model on different domains as suggested in the literature [Xie 2020, Chang 2019, Wu 2021]. We propose disentangled feature normalization (DFN) to effectively disentangle the feature learning for the datasets of different domains by replacing every group normalization (GN) layer in the feature extractor with two GN layers: one for the source domain images, *GN(S)*, and another for the target domain images, *GN(T)*. The updated model, called *km-rcnn++*, uses separate affine parameters at every normalization stage in the feature extractor for the source and the target domain images, efficiently normalizing the features of the two domains, see figure 6.2. The GN parameters for the target domain, *GN(T)*, are initialized from the source domain GN parameters, *GN(S)*, before the domain adaptation training.

The UDA approaches require weighing the losses differently for unlabeled and labeled images, usually to weigh more the unlabeled losses than the labeled ones to overcome the over-fitting to the labeled set. It can be easily performed if the underlying model is the same for the two domains: the usual case of the existing UDA approaches. However, with our improved design, the *km-rcnn++* model expects an input batch containing the first half of images from the source domain and the second half of images from the target domain. DFN therefore modifies the loss function described in equation 6.1 to compute and weigh the losses on the source and the target domain images differently. The input batch passes through the feature extractor, and the obtained features are divided into two halves corresponding to the source and the target domains. Each half then passes through the RPN network and the three heads to compute the separate RPN, bounding box, keypoint, and mask losses for source and the target domain images as given below.

$$\mathcal{L}_s = \sum_i \mathbb{L}_{cls}^{rpn}(f_i^l, y_i^l) + \mathbb{L}_{reg}^{rpn}(f_i^l, y_i^l) + \mathbb{L}_{cls}^{bbox}(f_i^l, y_i^l) + \mathbb{L}_{reg}^{bbox}(f_i^l, f_i^l) + \mathbb{L}_{ce}^{kp}(f_i^l, y_i^l) + \mathbb{L}_{bce}^{mask}(f_i^l, y_i^l) \tag{6.2}$$

$$\begin{aligned}\mathcal{L}_u = \sum_i & \mathbb{L}_{cls}^{rpn}(f_i^u, y_i^u) + \mathbb{L}_{reg}^{rpn}(f_i^u, y_i^u) + \mathbb{L}_{cls}^{bbox}(f_i^u, y_i^u) \\ & + \mathbb{L}_{reg}^{bbox}(f_i^u, f_i^u) + \mathbb{L}_{ce}^{kp}(f_i^u, y_i^u) + \mathbb{L}_{bce}^{mask}(f_i^u, y_i^u).\end{aligned} \tag{6.3}$$

Here, f_i^l and f_i^u correspond to the features of the labeled and unlabeled domain, respectively. The y_i^l corresponds to the source domain labeled ground-truth labels, and y_i^u correspond to the target domain pseudo labels. The following section explains the automatic generation of target domain pseudo labels y_i^u. The *km-rcnn++* in the inference mode uses only the GN layers corresponding to the target domain, thereby maintaining the same number of parameters and inference cost compared to *km-rcnn+*.

6.2.3 AdaptOR

Given a model, *km-rcnn++*, that can handle the datasets of different domains, we explain *AdaptOR*, our proposed method for unsupervised domain adaptation. We first explain *transformation equivariance constraints*, as needed to add explicit geometric constraints, and then the data augmentation pipeline, followed by the complete algorithm.

6.2.3.1 Transformation equivariance constraints

The state-of-the-art UDA or SSL approaches for image classification exploit the *transformation invariance* property on the unlabeled data, i.e., the classification labels remain unchanged irrespective of the transformation applied to the input image. However, the *invariance* property does not hold for the spatial localization tasks, and labels get changed with the viewpoint changes of the image due to geometric transforms, for example, resize and horizontal flip. But, these changes in the labels are *equivariant*

Algorithm 1 : *AdaptOR* algorithm to adapt a model trained on the labeled source domain dataset to the unlabeled target domain (operating room)

Inputs:

- Labeled dataset from the source domain $\mathcal{X} = \{x_i|y_i\}_{i=1}^{N_l}$, unlabeled dataset from target domain $\mathcal{U} = \{u_j\}_{i=1}^{N_u}$, $y_i = (\mathcal{P}_{bbox}, \mathcal{P}_{kp}, \mathcal{P}_{mask})$: ground-truth labels for the bounding box, keypoints, and mask for each person in the given labeled image.
- $p_t(y|x;\tilde{\phi})$: teacher model, $p_s(y|x;\phi)$: student model, $\tilde{\phi}$, ϕ: weights of the teacher and the student model respectively
- $\Gamma(p, \delta = \delta_{bbox}, \delta_{kp}, \delta_{mask})$: function to convert predictions (p) to pseudo labels using thresholds (δ) consisting of bounding box threshold δ_{bbox}, keypoint threshold δ_{kp}, and mask threshold δ_{mask}
- $\mathcal{T}_w(.)$: weak transform, $\mathcal{T}_s(.)$: strong transform
- $\mathcal{L}$: modified multi-task loss function as described in section 6.2.2.2 and equations 6.2 and 6.3, α: EMA decay rate, λ: unsupervised weight loss value, η: learning rate

Outputs: $\tilde{\phi}$: Final teacher model weights

1: **for all** $(\mathcal{X}_b, y_b, \mathcal{U}_b) \in (\mathcal{X}, \mathcal{U})$ **do** // sample a batch from the labeled and unlabeled dataset
2: $\mathcal{X}_w, y_w, \mathcal{U}_w = \mathcal{T}_w(\mathcal{X}_b, y_b, \mathcal{U}_b)$ // apply weak transform to the labeled and unlabeled batch to construct weakly augmented labeled $(\mathcal{X}_w, y_w)$ and unlabeled $(\mathcal{U}_w)$ batch
3: $\mathcal{X}_s, y_s, \mathcal{U}_s = \mathcal{T}_s(\mathcal{X}_b, y_b, \mathcal{U}_b)$ // apply strong transform to the labeled and unlabeled batch to construct strongly augmented labeled $(\mathcal{X}_s, y_s)$ and unlabeled $(\mathcal{U}_s)$ batch
4: $\tilde{y_s} = \Gamma(p_t(\mathcal{U}_w;\tilde{\phi}), \delta)$ // run the teacher model $p_t(y|x;\tilde{\phi})$ on the weakly augmented unlabeled batch $\mathcal{U}_w$, and convert the predictions into the pseudo labels $\tilde{y_s}$ using the thresholding function $\Gamma(p, \delta)$
5: $\bar{y_s} = \mathcal{T}_s(\mathcal{T}_w^{-1}(\tilde{y_s}))$ // apply the transform to convert the pseudo labels $\tilde{y_s}$ into the coordinates of strongly augmented unlabeled batch $(\mathcal{U}_s)$
6: $\mathcal{X}, y = concat(\mathcal{X}_w, \mathcal{X}_s, \mathcal{U}_s), concat(y_w, y_s, \bar{y_s})$ // concatenate the strongly augmented unlabeled batch with the weakly and strongly augmented labeled batch
7: $\mathcal{L}_s, \mathcal{L}_u = \mathcal{L}(p_s(\mathcal{X};\phi), y)$ // compute the loss using the multi-task loss function on the student model
8: $loss = \mathcal{L}_s + \lambda\mathcal{L}_u$ // add the supervised and the unsupervised losses
9: $\phi = SGD(\phi, \eta, \nabla_\phi(loss))$ // update the parameters of the student model ϕ using stochastic gradient descent with momentum
10: $\tilde{\phi} = \alpha\tilde{\phi} + (1-\alpha)\phi$ // update the parameters of teacher model $\tilde{\phi}$ using the exponential moving average
11: **end for**

to the applied transformations. Mathematically, if $\mathcal{F}(.)$ is a model that outputs the spatial localization labels for the input image I under transformation $\mathcal{T}$, we can minimize $\|\mathcal{F}(\mathcal{T}(I)) - \mathcal{T}(\mathcal{F}(I))\|$ under *transformation equivariance constraints*, i.e., the transformation $\mathcal{T}$ can be to used map the localization labels to the transformed image space. We use this property to provide the explicit geometric constraints on the unlabeled images. Additionally, specific to the human pose estimation under horizontal flipping transformation, we exploit the chirality transform [Yeh 2019] for the mapping of the human pose to the horizontally flipped image.

6.2.3.2 Data augmentations

Data Augmentations construct novel and realistic samples by computing stochastic transforms on the input data. The recent advancements in data augmentations have been the key to the performance boost in the supervised as well as SSL approaches [Cubuk 2019, Cubuk 2020, DeVries 2017]. We use two types of augmentations: *weak* and *strong*. The *weak* augmentations, $\mathcal{T}_w$, consist of random-flip and random-resize whereas *strong* augmentations, $\mathcal{T}_s$, consist of spatial augmentations from rand-augment [Cubuk 2020], random cut-out [DeVries 2017], random-flip, and random-resize, along with *strong-resize* augmentation to generate privacy-preserving low-resolution images. The *strong-resize* data augmentation down-sample and up-sample the input image with a random scaling factor chosen between 1x to 12x. Fig. 6.1 shows sample images from the OR at different downsampling scales.

6.2.3.3 Algorithm

Given the *weakly* augmented image, constructed using transformation $\mathcal{T}_w$, and the *strongly* augmented image, constructed using the transformation $\mathcal{T}_s$, our idea is to geometrically transform the pseudo labels - obtained from the model's predictions - of the *weakly* augmented image to the corresponding *strongly* augmented image. As the *weakly* and the *strongly* augmented images are generated using different geometric transformations with the pseudo labels being in the *weakly* augmented image coordinate system, we exploit *transformation equivariance constraints* to transform the pseudo labels by applying a transformation, $\mathcal{T}_s\mathcal{T}_w^{-1}$, to go from *weakly* augmented image space to the *strongly* augmented image space. The model is trained on the *strongly* augmented images with the transformed pseudo labels.
However, training the same model to generate and consume the pseudo labels may lead to unstable training. The *mean-teacher* [Tarvainen 2017] from semi-supervised learning has been proposed to stabilize the training using closely coupled *teacher* and a *student* model. We therefore adapt *mean-teacher* in our approach, where we use the *teacher* model to generate the pseudo labels on the *weakly* augmented image, and the *student* model to train on the corresponding *strongly* augmented image using the pseudo labels. As the source domain GN parameters, *GN(S)*, are trained under the direct supervision, we use *GN(S)* layers in the *teacher* model for the inference on the unlabeled target domain. The weights of the *teacher* and the *student* models are initialized from the same model, *kmrcnn++*. The weights of the *student* model are updated using the stochastic gradient descent based back-propagation, whereas the weights of the *teacher* model are updated using the exponential moving average (EMA) of the weights of the *student* model:

$$\tilde{\phi} = \alpha\tilde{\phi} + (1-\alpha)\phi,$$

where $\tilde{\phi}$ and ϕ are the weights of the teacher model and the student models, respectively, and α is a decay parameter. The EMA helps the *teacher* model to generate better

predictions due to its temporal ensembled weights from the *student* model, in turn improving the *student* model for better training. The detailed algorithm is explained in algorithm 1 and illustrated in figure 6.2.
Furthermore, we also test *AdaptOR* as an SSL approach, called *AdaptOR-SSL*, on a source domain dataset by making minimal changes. We use the *kmrcnn+* model, without disentangled feature normalization, as the images are coming from the same domain and do not concatenate the labeled and the unlabeled batches. The labeled and the unlabeled batches pass separately through the *kmrcnn+* model to calculate the separate losses on the labeled and the unlabeled data. The *AdaptOR-SSL* uses x%(x=1,2,5,10) of images from the source domain as the labeled dataset and the rest of the images as the unlabeled dataset.

6.3 Baselines

We first introduce several *self-training* based baselines that we have constructed for our joint person pose estimation and instance segmentation task by extending representative approaches. We extend pseudo-label [Lee 2013, Sohn 2020b], data-distillation [Radosavovic 2018], and ORPose, from the previous chapter, as our baselines approaches. We refer to the extended version of pseudo-label, data-distillation, and ORPose as KM-PL, KM-DDS, and KM-ORPose, respectively. The KM as a prefix signifies that these approaches have been extended for the joint pose (keypoint) estimation and instance (mask) segmentation tasks. The baselines approaches are two-stage approaches where the first stage generates the pseudo labels on the unlabeled data. The second stage jointly trains the model using the pseudo and the ground truth labels. *AdaptOR* on the other hand generates the pseudo labels on the unlabeled data *on-the-fly* during the training. For a fair comparison, we train all the baseline methods with the same training strategy, data augmentation pipeline, and *kmrcnn++* model. We give a brief overview of extended baseline approaches as follows.

6.3.1 KM-PL

We modify the pseudo-labeling [Lee 2013] approach to generate the pseudo labels on a single-scale image on the unlabeled target domain data. The authors in [Sohn 2020b] recently use a similar approach with advanced data augmentations for the object detection task.

6.3.2 KM-DDS

KM-DDS [Radosavovic 2018] is also a pseudo-labeling approach, but instead of generating pseudo labels on a single scale, it aggregates the labels from multiple scales with random horizontal flipping transformations. Authors use the approach for multi-class object detection and human pose estimation. We further extend it to generate pseudo labels for the masks. Similar to the authors, we use scaling and random horizontal flipping transformations on nine predefined image sizes ranging from 400 to

Table 6.1: An overview of the source and the target domain datasets used in this work.

Dataset	**type**	**# images**	**# instances**
Source domain labeled dataset			
COCO	train	57,000	150,000
COCO-val	test	5,000	10,777
Target domain unlabelled datasets			
MVOR	train	80,000	-
MVOR+	test	2,196	5,091
TUM-OR	train	1,500	-
TUM-OR-test	test	2,400	11,611

1200 pixels with a step size of 100. Here, the image size corresponds to the shorter side of the image; the size of the longer side of the image is computed by maintaining the same aspect ratio.

6.3.3 KM-ORPose

KM-ORPose uses the *teacher-student* learning paradigm for the domain adaptation in the OR for joint person detection and 2D/3D human pose estimation as described in previous chapter. It combines the knowledge-distillation [Hinton 2015, Zhang 2019a] - using complex three-stage models - along with data-distillation [Radosavovic 2018] to generate accurate pseudo labels. In the first stage, it uses cascade-mask-rcnn [Cai 2019b] with the deformable convolution [Dai 2017] based resnext-152 [Xie 2017] to generate the person bounding boxes. We use the same network to get the pseudo masks as well. In the second stage, it uses the HRNet-w48 model (384x288 input size) [Sun 2019] to get the pseudo labels for the poses. KM-ORPose is a strong baseline as it uses a complex multi-stage teacher model to generate accurate pseudo labels for the training.

6.4 Experiments

6.4.1 Datasets and Evaluation Metrics

We use COCO [Lin 2014] as source domain dataset. It contains 57k images and 150k person bounding boxes along with a segmentation mask and 17 body keypoints. The test dataset of COCO, called *COCO-val*, contain 5k images with 10777 person instances.

We train our approach on an unlabelled training dataset of MVOR, called *MVOR-unlabeled* (see section 3.2.1), and TUM-OR containing 80k and 1.5k images, respectively. We evaluate our approach on the two target domain OR datasets: *MVOR+* as described in section 3.2.2.2 and *TUM-OR-test* [Belagiannis 2016]. The original *TUM-OR-test* consists of only the upper-body bounding boxes with six common COCO keypoints. These annotations are not suitable for our evaluation purpose; hence we annotate the *TUM-OR-test* using a semi-automatic approach. We first use a state-of-the-art person detector [Cai 2019b] to get the person bounding boxes and manually correct all the bounding boxes. We then run the HRNet model [Sun 2019] on

all the corrected bounding boxes to get the poses. The predicted poses are corrected using the keypoint annotation tool [2]. An overview of the datasets used in this work is shown in the Table 6.1

The image sizes of *MVOR+* and *TUM-OR-test* datasets are 640x480 and 1280x720, respectively. We also conduct experiments with downsampled images using the scaling factors 8x, 10x, and 12x, yielding images of size 80x64, 64x48, and 53x40 for the *MVOR+* dataset and 160x90, 128x72, and 107x60 for the *TUM-OR-test* dataset.

We use AP^{bb}_{person} for person bounding box evaluation, AP^{bb}_{person} for 2D keypoint evaluation, and $AP^{bb\ (from\ mask)}_{person}$ for instance segmentation, respectively, as described in section 3.3.3.

6.4.2 Experiments

6.4.2.1 Source domain fully supervised training

The models are trained on the source domain COCO dataset in a fully supervised manner for three experiments: supervised ImageNet initialization with Frozen batch normalization (BN) [He 2016], self-supervised MOCO-v2 initialization [Chen 2020, He 2020] with Cross-GPU BN [Peng 2018], and self-supervised MOCO-v2 initialization [Chen 2020, He 2020] with group normalization (GN) [Wu 2018].

The goal of these experiments is to obtain one suitable *source-only* baseline as an initialization model for the UDA experiments. The last model with self-supervised MOCO-v2 initialization and GN, called *kmrcnn+*, is further used in the SSL experiments and extended in UDA experiments.

6.4.2.2 AdaptOR: unsupervised domain adaptation (UDA) on target domains

The UDA experiments on source domain COCO datasets and target domains MVOR and TUM-OR datasets are conducted to train the *kmrcnn++* model for eight sets of experiments. The first four experiments are for the target domain MVOR and the last four for TUM-OR. For each target domain, the first three experiments train the *kmrcnn++* model on three constructed baseline methods: KM-PL, KM-DDS, and KM-ORPose, respectively, and the fourth experiment trains the *kmrcnn++* model on our *AdaptOR* method. Eleven ablation experiments are conducted with the source domain COCO dataset and the target domain MVOR dataset: the first experiment evaluates the contribution of disentangled feature normalization, the next five different types of strong augmentations, and the last five different unsupervised loss weights loss values λ.

6.4.2.3 AdaptOR-SSL: semi-supervised learning (SSL) on source-domain

The SSL experiments on the source domain COCO dataset are conducted for four experiments where we train the *kmrcnn+* model using 1%, 2%, 5%, and 10% of COCO

[2] https://github.com/visipedia/annotation_tools

dataset as the labeled set and the rest of the data as the unlabeled set. The *kmrcnn+* model uses the regular GN layers instead of disentangled feature normalization layers. We use the same labeled and unlabeled images and training iterations as used by Unbiased-teacher [Liu 2021b], the current state-of-the-art in SSL for object detection.

6.4.2.4 Domain adaptation on AdaptOR-SSL model

AdaptOR assumes it has access to all the source-domain labels in the previous experiments. We conduct a final experiment to see how *AdaptOR* performs when initialized from a source-domain model trained with less source domain data. We take a *AdaptOR-SSL* model trained using 10% labeled and 90% unlabeled source domain data and use it to initialize *AdaptOR*.

6.4.3 Implementation details

The source domain fully supervised training experiments, explained in section 6.4.2.1, are conducted with batch size 16 and learning rate 0.02 for 270k iterations with multi-step (210k and 250k) learning rate decay on eight V100 GPUs.

The *AdaptOR* and the *AdaptOR-on-AdaptOR-SSL* experiments explained in section 6.4.2.2, 6.4.2.4, respectively, are conducted on four V100 GPUs with a labeled and unlabeled batch size of eight (four images/GPU) and a learning rate of 0.001. The experiments are conducted for 65k iterations for the MVOR dataset and 10k iterations for the TUM-OR dataset. Finally, the *AdaptOR-SSL* experiments explained in 6.4.2.3 are conducted on four V100 GPUs following the linear learning rate scaling rule [Goyal 2017].

The spatial augmentations from rand-augment [Cubuk 2020] consist of "inversion", "auto-contrast", "posterize", "equalize", "solarize", "contrast-variation", "color-jittering", "sharpness-variations", and "brightness-variations" implemented using a python image library[3]. The random cut-out [DeVries 2017] augmentation places square boxes of random sizes chosen between 40 to 80 pixels at random locations in the image. The random-resize operation for the *weakly* and *strongly* augmented images resize the image to a size randomly sampled from 600 to 800 pixels for SSL experiments following [He 2017]. For the UDA experiments, we choose the random-resize range from 480 to 800 pixels to provide more size variability in the data augmentation and match the original size of the MVOR dataset (640x480). The image size corresponds to the shorter side of the image.

We use a detectron2 framework [Wu 2019a] to run all the experiments with automatic mixed precision (AMP) [Micikevicius 2017]. We use bounding box threshold $\delta_{bbox} = 0.7$, keypoint threshold $\delta_{kp} = 0.1$, mask threshold $\delta_{mask} = 0.5$, EMA decay rate $\alpha = 0.9996$, unsupervised loss weight $\lambda = 3.0$ for *AdaptOR*, and $\lambda = 2.0$ for *AdaptOR-SSL*.

[3] https://github.com/jizongFox/pytorch-randaugment

Table 6.2: Results on the source domain *COCO-val* dataset with 100% labeled supervision. The *kmrcnn+* model using GN [Wu 2018] and initialized using self-supervised MoCo-v2 approach [Chen 2020, He 2020] perform equally well with the model using Cross-GPU BN [Peng 2018] but using less training time. The first row results for the *kmrcnn* model is obtained from the paper [He 2017]. Rest of the results correspond to the models that we train. Inference is performed on a single-scale of 800 pixels following [He 2017]. Automatic mixed precision (AMP) uses single- and half-precision (32 bits and 16 bits) floating operation to speed up the training while trying to maintain single-precision (32 bits) model accuracy.

Model	initialization	Normalization	AMP	≈ Training-time	AP^{bb}_{person}	AP^{kp}_{person}	AP^{mask}_{person}
kmrcnn	Supervised-Imagenet	Frozen BN	✗	32 hours	52.0	64.7	45.1
kmrcnn	Supervised-Imagenet	Frozen BN	✓	16 hours	56.4	65.7	49.1
kmrcnn	MoCo-v2	Cross-GPU BN	✓	22 hours	57.5	66.6	49.8
kmrcnn+	MoCo-v2	GN	✓	18 hours	57.5	66.2	49.9

6.5 Results

6.5.1 Source domain fully supervised training

Table 6.2 shows the results of *kmrcnn* and *kmrcnn+* models trained on the source domain COCO dataset. The *kmrcnn* trained using self-supervised MoCo-v2 weights with Cross-GPU BN [Peng 2018] obtains improvement of approximately 1% in all the three metrics compared to supervised ImageNet weights using frozen BN. The *kmrcnn+* using GN performs equally well but with less training time. The *kmrcnn+* model is therefore further used in the SSL experiments and extended in UDA experiments.

6.5.2 AdaptOR: unsupervised domain adaptation (UDA) on target domains

Table 6.3 and figure 6.3 show the result of our unsupervised domain adaptation experiments using *AdaptOR*. The first and the second half in table 6.3 show the results for *MVOR+* and *TUM-OR-test* datasets, respectively. We evaluate the models at four downsampling scales (1x, 8x, 10x, and 12x). As the model is trained on unlabeled image sizes from 480 to 800 pixels (shorter side), we evaluate the model on nine target resolutions (480, 520, 560, 600, 640, 680, 720, 760, and 800), i.e., for a given downsampling scale, we down-sample the image with the scale and up-sample it to the given target resolution. The target resolution also corresponds to the shorter size of the image to maintain the aspect ratio. We use bilinear interpolation for the downsampling and up-sampling. The results in Table 6.3 show the mean and standard deviation of the results computed on all the target resolutions for bounding box detection AP^{bb}_{person}, pose estimation AP^{kp}_{person}, and instance segmentation AP^{bb}_{person} *(from mask)* on a given downsampling scale.

The first row shows the *source-only* results for the *kmrcnn+* model trained on source domain images and evaluated on the target domain. The significant decrease in the low-resolution results of the *kmrcnn+* is likely because such heavily downsampled images

[4] https://github.com/matteorr/coco-analyze

Table 6.3: Results for the baseline approaches and *AdaptOR*. We see improvements in all three metrics on both the target domain datasets, especially on the low-resolution images making the proposed approach suitable for the deployment inside the privacy-sensitive OR environment. The *source-only* results correspond to the model trained on the labeled source domain without any training on the target domain images. The KM-PL, KM-DDS, and KM-ORPose are strong baselines proposed in this work.

Methods	***MVOR+***				***TUM-OR-test***			
	1x	**8x**	**10x**	**12x**	**1x**	**8x**	**10x**	**12x**
	$\mathbf{AP^{bb}_{person}}$ **(mean±std)**							
source-only	56.61±0.34	40.42±2.17	34.87±2.47	29.61±2.69	68.61±1.54	41.84±2.33	31.08±2.83	24.00±2.90
KM-PL	60.21±0.51	57.14±0.34	55.88±0.39	54.26±0.41	72.28±1.51	65.44±1.45	62.84±1.02	62.42±1.55
KM-DDS	60.79±0.47	57.88±0.39	56.74±0.37	55.12±0.45	72.51±1.45	65.98±1.18	63.87±0.99	62.68±1.32
KM-ORPose	58.88±0.69	55.14±0.56	53.81±0.52	51.96±0.47	69.73±1.22	63.46±0.93	60.71±0.73	60.14±0.94
AdaptOR	**61.41±0.40**	**59.48±0.35**	**58.55±0.36**	**57.33±0.43**	**72.75±0.88**	**67.33±0.78**	**65.53±0.57**	**65.65±0.66**
	$\mathbf{AP^{kp}_{person}}$ **(mean±std)**							
source-only	50.55±0.39	23.99±2.25	16.86±2.16	11.31±1.91	65.60±4.55	27.21±1.49	19.41±1.86	13.18±1.81
KM-PL	58.72±0.44	55.19±0.43	52.81±0.55	49.53±0.46	77.49±1.87	67.57±1.03	63.46±0.89	58.24±1.05
KM-DDS	59.83±0.40	55.60±0.49	53.16±0.48	50.02±0.46	78.39±1.76	69.24±1.07	65.29±0.93	60.56±1.21
KM-ORPose	**62.50±0.53**	57.18±0.60	54.59±0.59	51.24±0.47	**80.49±1.74**	69.90±1.03	65.64±0.94	60.67±0.73
AdaptOR	60.86±0.38	**57.35±0.61**	**55.42±0.66**	**52.60±0.60**	77.84±1.24	**70.65±1.04**	**67.36±0.96**	**63.27±1.21**
	$\mathbf{AP^{bb\ (from\ mask)}_{person}}$ **(mean±std)**							
source-only	54.95±0.37	37.98±2.21	32.58±2.37	27.56±2.48	69.33±1.46	40.38±2.30	30.11±2.79	22.97±2.93
KM-PL	56.50±0.60	54.06±0.44	52.90±0.48	51.33±0.46	71.93±1.34	65.43±1.46	63.16±0.89	62.67±1.11
KM-DDS	57.12±0.47	54.76±0.50	53.78±0.49	52.06±0.67	71.99±1.18	65.96±1.07	64.02±0.70	63.01±1.02
KM-ORPose	55.46±0.76	52.37±0.62	51.23±0.55	49.34±0.46	68.05±1.13	61.15±1.09	58.53±0.86	57.89±1.00
AdaptOR	**59.34±0.40**	**57.44±0.42**	**56.62±0.41**	**55.39±0.51**	**72.13±0.91**	**66.55±0.80**	**65.04±0.52**	**65.15±0.65**

are not present in the source domain. The improved result for the KM-DDS approach compared to KM-PL shows the effects of generating pseudo labels using the multi-scale and flipping transformation. The bounding box and segmentation results for the KM-ORPose are slightly worse than the KM-PL and KM-DDS. It may be because KM-ORPose uses a state-of-the-art object detector trained on all the 80 class categories from COCO whereas, KM-PL and KM-DDS use the model trained specifically for the person class. The *AdaptOR* performs significantly better compared to baseline approaches, especially on the low-resolution at different target resolutions, see figure 6.3, suggesting the potential of our approach for low-resolution images in the privacy-sensitive OR environment. We observe a slight decrease in the accuracy for AP^{kp}_{person} metric on original size, likely due to the use of the multi-stage complex teacher model to generate the pseudo poses. Instead, our approach improves the given model in a model agnostic way without relying on an external teacher model to generate the pseudo labels. We also plot the results at individual scales in the figure 6.3. The figure 6.4 and 6.5 show qualitative results comparing our approach with the baseline approaches.

We further analyze the impact of different localization errors at the keypoint level before and after the domain adaptation using an approach described in [Ruggero Ronchi 2017]. As shown in Fig. 6.7, after domain adaptation, our approach correctly detects more

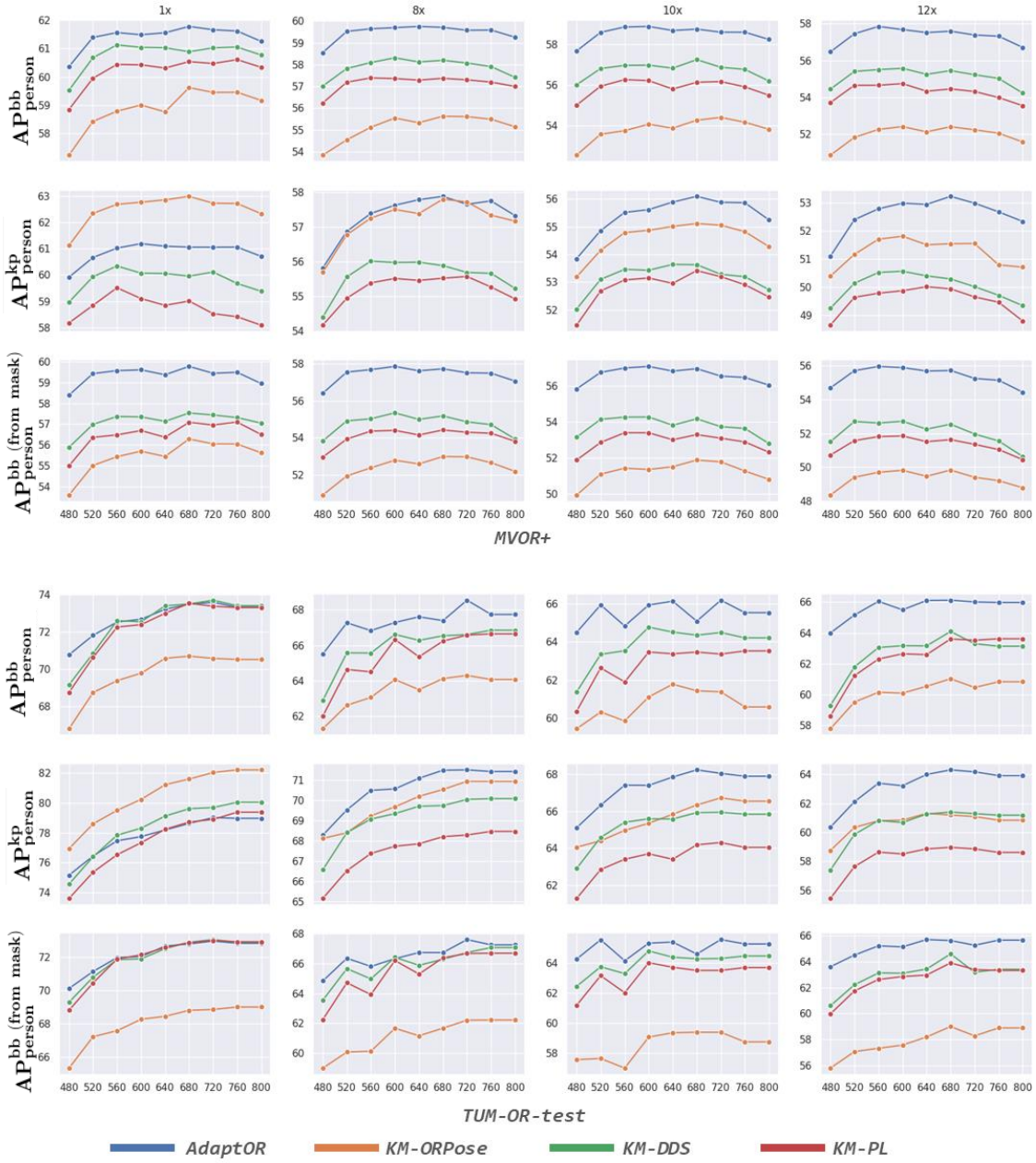


Figure 6.3: Bounding box detection AP^{bb}_{person}, pose estimation AP^{kp}_{person}, and instance segmentation AP^{bb}_{person} *(from mask)* results for unsupervised domain adaptation experiments on four downsampling scales (1x, 8x, 10x, and 12x) and nine target resolution (480, 520, 560, 600, 640, 680, 720, 760, and 800) corresponding to the shorter side of the image for *MVOR+* and *TUM-OR-test* datasets. We see an increase in the accuracy with the increase in target resolution for the *TUM-OR-test* dataset. We also observe an increase in accuracy for the *MVOR+* dataset but only up to around 680 pixels.

keypoints while reducing the impact of different localization errors. Additional qualitative results for the UDA experiments on *MVOR+* and *TUM-OR-test* are

Figure 6.4: Qualitative results for bounding box detection, pose estimation, and instance segmentation on a sample *MVOR+* image for the baseline approaches and *AdaptOR*. Results are displayed on the for original image and corresponding downsampled images with downsampling factor 8 and 12. The red arrows show either missed detections or localization errors. Localization errors are noticeable on the low-resolution images.

presented in the supplementary video[5]

[5] https://youtu.be/gqwPu9-nfGs

Figure 6.5: Qualitative results for bounding box detection, pose estimation, and instance segmentation on a sample *TUM-OR-test* image for the baseline approaches and *AdaptOR*.

6.5.2.1 Ablation experiments

Disentangled feature normalization

Fig. 6.8 shows t-sne feature visualization [Van der Maaten 2008] of the *layer5* resnet features of the backbone model illustrating the appropriate segregation of the features after the domain adaptation. We also conduct experiments to quantify the use of two separate GN layers, *GN(S)* and *GN(T)*, in the feature extractor for domain-specific normalization compared to either a single GN layer or a single frozen BN layer. The first row in Table 6.4 shows the results for the *krcnn* [He 2017, Wu 2019b] model using frozen BN [He 2016] layers for joint bounding box detection and pose estimation. We take the source domain COCO trained weights from detectron2 [Wu 2019a] library and train it on the MVOR dataset. The second row shows the results for the *kmrcnn+* model using a single GN layer for both domains. We also evaluate *kmrcnn++* where we use the GN

1% supervision 2% supervision

5% supervision 100% supervision

Figure 6.6: Qualitative results on a sample image from *COCO-val* dataset with x%(x=1,2,5,100) of labeled supervision. We use the *AdaptOR-SSL* for 1%, and 5% labeled supervision with the rest of the data as the unlabeled data. We see comparable qualitative results with 1% of labeled supervision to 100% of labeled supervision. The red arrows show either missed detections or localization errors.

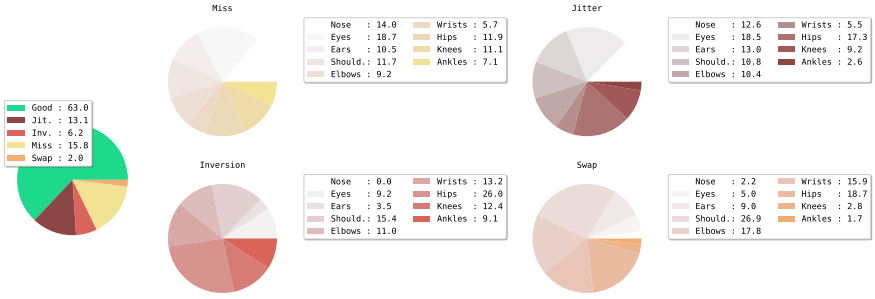
layers corresponding to source domain *GN(S)* to evaluate on the target domain (*kmrcnn++ GN(S)*). We obtain significantly better results by using our design of the two separate GN layers for feature normalization.

Components of AdaptOR

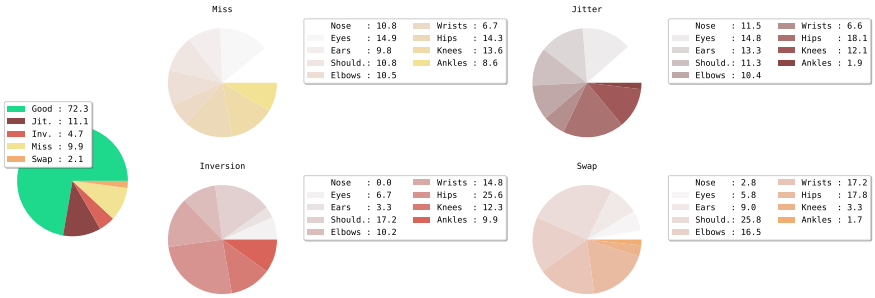
Table 6.5 shows the ablation experiments to see the effect of using different types of augmentations on the strongly transformed images used by the student model during training. The results show that the *strong-resize* augmentations are needed to adapt the model to the low-resolution OR images. The geometric transform exploiting the *transformation equivariance constraints* significantly improves the results, especially for the pose estimation task, where we also utilize the chirality transforms to map the flipped keypoints to the horizontally flipped image. The results are further improved using the random-augment and random-cut augmentations.

Effect of unsupervised loss weight (λ) values

Unsupervised loss weight (λ) controls the proportion of the total loss attributed to the unsupervised loss for the target domain. As the aim is to adapt the model to the target domain, higher value of λ generally leads to better performance. Fig. 6.9 shows the ablation results for different values of unsupervised loss weight (λ). We observe that the increase in the λ increases the accuracy; however, it starts to decrease after the λ value of 4.0.



(a) *source-only*



(b) *AdaptOR*

Figure 6.7: Localization errors at individual keypoint level for the pose estimation task before and after the domain adaptation. "Jitter", "Inversion", "Swap", and "Miss" are various localization errors defined in [Ruggero Ronchi 2017]: "Jitter" error is the error in predicted keypoint w.r.t close proximity of the correct ground truth, "Inversion" error is due to the right-left swap of the body part, "Swap" is the error in assigning predicted keypoint to a wrong person, and "Miss" error is due to completely missing the correct ground truth location. We use the author's code repository [Ruggero Ronchi 2017][4] for plotting the results.

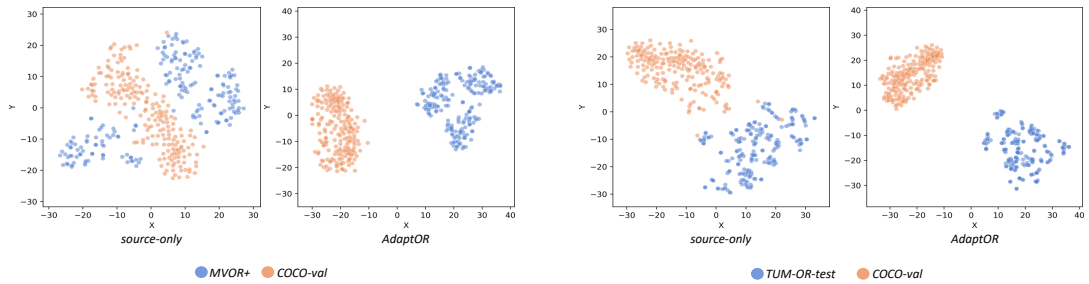


Figure 6.8: t-sne feature visualization [Van der Maaten 2008] of the *layer5* resnet features of the backbone model on random 200 images of the source and the target domain test datasets. The *source-only* model uses only the *GN(S)* layers whereas the *AdapOR* uses separate *GN(S)* and *GN(T)* layers for the source and the target domain images, respectively. The *AdapOR* model appropriately segregates the source and the target domain image features from the two domains helping in improving the domain adaptation for the downstream heads.

6.5.3 AdaptOR-SSL: semi-supervised learning (SSL) on source-domain

Table 6.6 shows the results of SSL experiments using *AdaptOR-SSL* on the COCO dataset with 1%, 2%, 5%, and 10% labeled supervision. The results with 100% labeled

Table 6.4: Ablation study comparing the *kmrcnn++* model using the two GN layer-based design for feature normalization with the *kmrcnn+* that uses only a single layer. We also compare it with a *krcnn* model using single frozen BN, and *kmrcnn++ GN(S)*, the same *kmrcnn++* model but using the GN layers corresponding to the source domain.

Models	***MVOR+***			
	1x	**8x**	**10x**	**12x**
	$\mathbf{AP^{bb}_{person}}$ **(mean±std)**			
krcnn	59.00±0.35	56.78±0.37	55.87±0.34	54.43±0.36
kmrcnn+	60.71±0.16	58.75±0.33	58.03±0.31	56.97±0.39
kmrcnn++ GN(S)	59.64±0.46	55.86±0.48	53.84±0.64	51.61±0.74
kmrcnn++	**61.41±0.40**	**59.48±0.35**	**58.55±0.36**	**57.33±0.43**
	$\mathbf{AP^{kp}_{person}}$ **(mean±std)**			
krcnn	57.96±0.32	55.48±0.62	53.34±0.55	50.50±0.44
kmrcnn+	47.15±0.30	45.27±0.44	43.89±0.44	42.01±0.44
kmrcnn++ GN(S)	58.64±0.40	52.37±0.41	49.51±0.46	46.08±0.51
kmrcnn++	**60.86±0.38**	**57.35±0.61**	**55.42±0.66**	**52.60±0.60**
	$\mathbf{AP^{bb\ (from\ mask)}_{person}}$ **(mean±std)**			
krcnn	-	-	-	-
kmrcnn+	55.18±0.25	53.85±0.42	53.28±0.5	52.44±0.58
kmrcnn++ GN(S)	58.22±0.46	54.77±0.62	53.06±0.67	50.70±0.72
kmrcnn++	**59.34±0.40**	**57.44±0.42**	**56.62±0.41**	**55.39±0.51**

copy.png copy.png

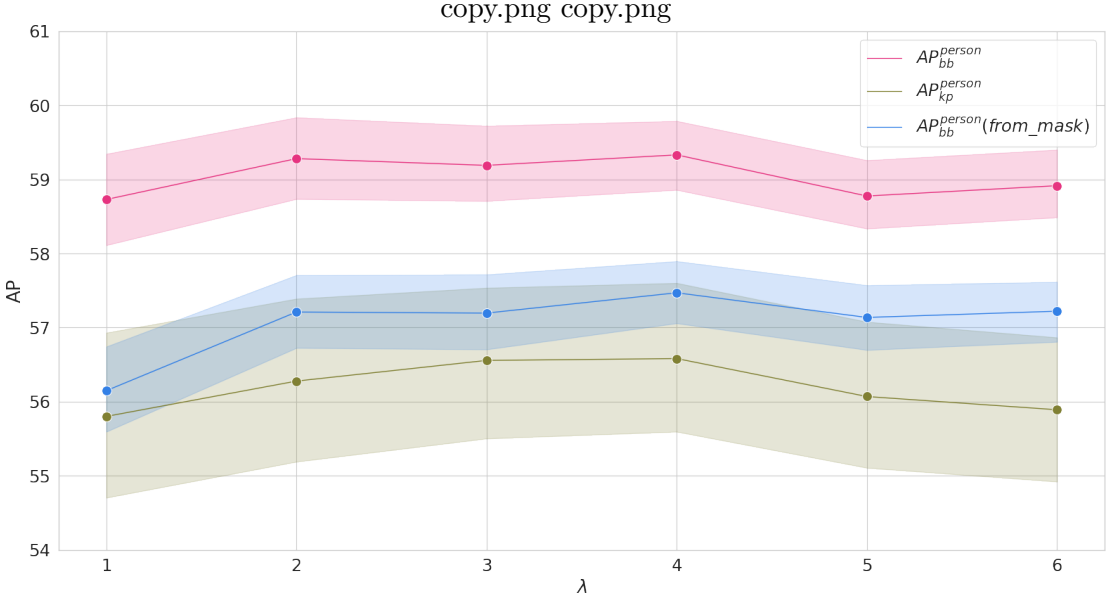


Figure 6.9: Results for different values of unsupervised loss weight (λ) on the *MVOR+* dataset. Results show the mean and confidence interval computed using different downsampling scales (1x, 8x, 10x, and 12x) and target resolutions (480, 520, 560, 600, 640, 680, 720, 760, and 800).

supervision are presented in Table 6.2. The first two rows in Table 6.6 show the results of two fully supervised baselines: *supervised* and *supervised++*. The *supervised* baseline uses random-resize and random-flip augmentations as used [He 2017], whereas the

Table 6.5: Ablation study quantifying the different augmentations on the strongly transformed image used by the student model for the training. Here, sr: *strong-resize*, ra: random-augment, rc: random-cut, and geom: geometric transformations consisting of random-resize and random-flip.

				MVOR+			
sr	ra	rc	geom	**1x**	**8x**	**10x**	**12x**
				$\mathbf{AP^{bb}_{person}}$ **(mean±std)**			
		Baseline		56.61±0.34	40.42±2.17	34.87±2.47	29.61±2.69
✗	✗	✗	✗	58.06±0.28	45.14±1.70	40.19±2.09	35.45±2.28
✓	✗	✗	✗	58.34±0.34	58.03±0.31	57.25±0.33	55.97±0.33
✓	✗	✗	✓	59.64±0.34	58.74±0.30	58.01±0.36	56.80±0.32
✓	✓	✗	✗	58.43±0.31	57.72±0.31	56.99±0.33	55.58±0.29
✓	✓	✓	✗	59.79±0.54	58.38±0.44	57.48±0.45	56.21±0.46
✓	✓	✓	✓	**61.41±0.40**	**59.48±0.35**	**58.55±0.36**	**57.33±0.43**
				$\mathbf{AP^{kp}_{person}}$ **(mean±std)**			
		Baseline		50.55±0.39	23.99±2.25	16.86±2.16	11.31±1.91
✗	✗	✗	✗	52.32±0.30	31.33±1.56	24.48±2.07	18.19±1.97
✓	✗	✗	✗	54.22±0.39	53.53±0.63	51.65±0.58	49.13±0.58
✓	✗	✗	✓	57.07±0.31	55.41±0.62	53.68±0.55	51.19±0.48
✓	✓	✗	✗	54.51±0.24	52.67±0.62	50.74±0.68	47.97±0.50
✓	✓	✓	✗	57.44±0.37	54.73±0.47	52.64±0.47	49.96±0.49
✓	✓	✓	✓	**60.86±0.38**	**57.35±0.61**	**55.42±0.66**	**52.60±0.60**
				$\mathbf{AP^{bb\ (from\ mask)}_{person}}$ **(mean±std)**			
		Baseline		54.95±0.37	37.98±2.21	32.58±2.37	27.56±2.48
✗	✗	✗	✗	56.08±0.32	42.12±1.78	37.19±2.13	32.56±2.27
✓	✗	✗	✗	55.81±0.38	55.66±0.46	54.94±0.43	53.73±0.51
✓	✗	✗	✓	57.14±0.35	56.52±0.38	55.84±0.42	54.62±0.41
✓	✓	✗	✗	56.06±0.32	55.50±0.33	54.70±0.41	53.30±0.40
✓	✓	✓	✗	57.58±0.50	56.34±0.45	55.48±0.50	54.21±0.62
✓	✓	✓	✓	**59.34±0.40**	**57.44±0.42**	**56.62±0.41**	**55.39±0.51**

supervised++ uses the our data augmentation pipeline containing *weakly* and *strongly* augmented labeled images. We observe significant improvement in the results by utilizing our data augmentation pipeline. We also compare our bounding box detection results with the current state-of-the-art SSL approach for object detection, Unbiased-teacher [Liu 2021b]: a multi-class object bounding box detection approach using *self-training* and *mean-teacher* based SSL approach. Different from ours, it uses fully supervised ImageNet weights for initialization and does not exploit the *transformation equivariance constraints* using geometric augmentations. As the Unbiased-teacher performs bounding box detection on 80 COCO classes, we compare our results with their person category results AP^{bb}_{person} from the model obtained from their GitHub repository[6]. We observe significant improvement in results attributed to our initialization using the self-supervised method, feature normalization using GN,

[6] https://github.com/facebookresearch/unbiased-teacher

Table 6.6: Results for *AdaptOR-SSL* on *COCO-val* dataset under the semi-supervised learning setting with x%(x=1,2,5,10) of labeled supervision. We compare it with the fully supervised baselines trained on the same labeled data without using any unlabeled data. The *supervised* baseline uses only the random resize and random-flip data augmentations as used in [He 2017] whereas *supervised++* uses the same data augmentation pipeline as in *AdaptOR-SSL* containing *weakly* and *strongly* augmented labeled images. We also compare it with the current state-of-the-art SSL object detector Unbiased-Teacher [Liu 2021b] for the person bounding box detection task. The inference is performed on a single scale of 800 pixels (shorter side) following the same settings as used in [He 2017, Liu 2021b].

Methods	$\mathbf{AP^{bb}_{person}}$				$\mathbf{AP^{kp}_{person}}$				$\mathbf{AP^{mask}_{person}}$			
	1%	2%	5%	10%	1%	2%	5%	10%	1%	2%	5%	10%
supervised	22.09	28.43	34.52	37.88	15.91	23.58	30.96	37.77	18.13	23.93	29.34	33.47
supervised++	28.59	34.27	41.18	43.60	25.78	32.14	41.45	46.51	24.18	29.14	35.40	37.83
Unbiased-Teacher	39.18	40.76	43.72	46.64	-	-	-	-	-	-	-	-
AdaptOR-SSL	**42.57**	**45.37**	**49.90**	**52.70**	**38.22**	**44.08**	**49.79**	**56.65**	**36.06**	**38.96**	**43.10**	**45.46**

exploitation of the geometric constraints on the unlabeled data, and single class training for person category exploiting mask and the keypoint annotations. Fig. 6.6 shows the qualitative results from the models trained with 1%, 2%, 5% and 100% labels. We also show the qualitative results in the supplementary video on some YouTube videos and observe comparable qualitative results with 1% of labeled supervision w.r.t 100% of labeled supervision.

6.5.4 Domain adaptation on AdaptOR-SSL model

Table 6.7 shows results when we evaluate *AdaptOR-SSL* models trained using 1%, 2%, 5%, and 10% source domain labels to our *MVOR+* target domain. We observe a significant decrease in the results (see "Before UDA" results in the Table 6.7). As a final experiment, we initialize our *AdaptOR* approach with *AdaptOR-SSL* model trained using 10% source domain labels. We observe an increase in the performance after the domain adaptation. However, there still exists a gap of around 4% for AP^{bb}_{person} and $AP^{bb\ (from\ mask)}_{person}$, and 12% for AP^{kp}_{person}. These results show the need to develop effective domain adaptation approaches in the presence of limited source domain labels.

6.6 Conclusion

Manual annotations, especially for spatial localization tasks, are considered the main bottleneck in the design of AI systems. With advances in digital technology providing a wide variety of visual signals, the modern OR has started to use AI to develop next-generation smart assistance systems. However, the progress is hindered due to the cost and privacy concerns for obtaining manual annotations. In this work, we tackle the joint person pose estimation and instance segmentation task needed to analyze OR activities and propose an unsupervised domain adaptation approach to adapt a model trained on a labeled source domain to an unlabeled target domain. We propose a new

Table 6.7: Performance comparison when applying *AdaptOR-SSL* models trained with 1%, 2%, 5%, and 10% source domain labels to the target domain of *MVOR+* (see "Before UDA" results). When we apply the *AdaptOR* approach on the *AdaptOR-SSL* model (trained using 10% source domain labels), we observe an improvement in the performance (see "After UDA" results). Results corresponding to 100% source domain labeled supervision in "Before UDA" and "After UDA" are obtained from Table 6.2 and 6.3, respectively.

models	***MVOR+***			
	1x	**8x**	**10x**	**12x**
	AP^{bb}_{person} (mean±std)			
Before UDA				
1%	48.33±0.67	39.89±1.97	34.46±2.44	28.63±3.25
2%	48.28±0.64	41.12±2.00	35.93±2.16	30.51±2.54
5%	51.27±0.48	43.11±2.08	37.95±2.14	31.75±2.72
10%	53.95±0.65	44.74±1.92	39.83±2.03	34.13±2.60
100%	56.61±0.34	40.42±2.17	34.87±2.47	29.61±2.69
After UDA				
10%	57.58±0.56	55.80±0.60	54.70±0.51	53.44±0.38
100%	61.41±0.40	59.48±0.35	58.55±0.36	57.33±0.43
	AP^{kp}_{person} (mean±std)			
Before UDA				
1%	25.28±1.06	16.64±1.17	12.72±1.90	08.34±1.94
2%	30.16±0.58	21.44±1.91	16.28±2.22	11.35±2.39
5%	37.09±0.30	25.93±2.22	20.12±2.38	13.84±2.50
10%	41.51±0.58	28.57±1.88	22.57±2.15	16.17±2.38
100%	50.55±0.39	23.99±2.25	16.86±2.16	11.31±1.91
After UDA				
10%	48.52±0.50	45.73±0.56	43.74±0.47	40.90±0.44
100%	60.86±0.38	57.35±0.61	55.42±0.66	52.60±0.60
	$AP^{bb\ (from\ mask)}_{person}$ (mean±std)			
Before UDA				
1%	47.54±0.78	38.37±2.32	32.44±2.73	26.32±3.42
2%	47.96±0.90	39.32±2.29	33.54±2.30	27.87±2.44
5%	50.55±0.74	41.09±2.30	35.68±2.16	29.45±2.69
10%	52.79±0.69	42.63±2.17	37.18±2.10	31.41±2.60
100%	54.95±0.37	37.98±2.21	32.58±2.37	27.56±2.48
After UDA				
10%	55.60±0.52	54.07±0.58	53.00±0.49	51.55±0.35
100%	59.34±0.40	57.44±0.42	56.62±0.41	55.39±0.51

self-training based framework with advanced data augmentations to generate pseudo labels for the unlabeled target domain. The high-quality effectiveness in the pseudo labels is ensured by applying explicit geometric constraints of the different augmentations on the unlabeled input image. We also introduce disentangled feature normalization for the statistically different source and the target domains and use the *mean-teacher* paradigm to stabilize the training. Evaluation of the method on the two

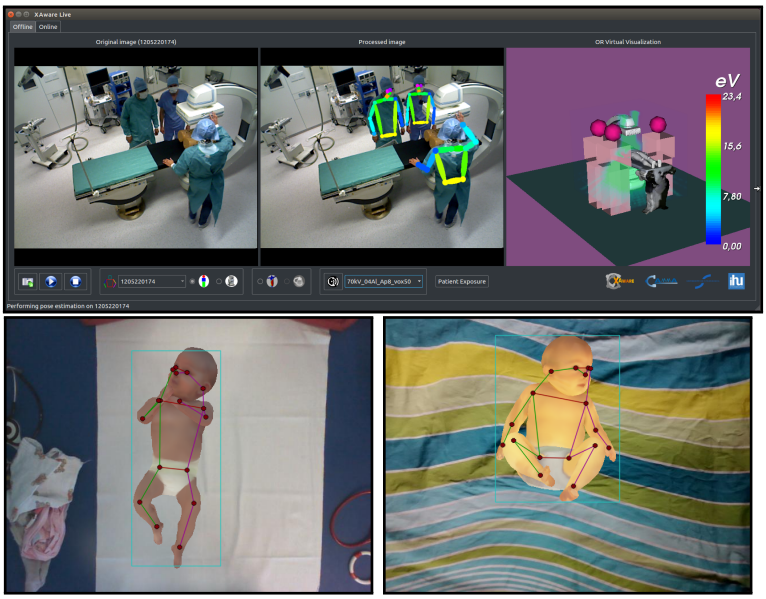
target domain datasets, *MVOR+* and *TUM-OR-test*, with extensive ablation studies, show the effectiveness of our approach. We further demonstrate that the proposed approach can effectively be adapted to the low-resolution images of the target domain, as needed to ensure OR privacy, even up to a downsampling factor of 12x. Finally, we illustrate the generality of our approach as the SSL method on the large-scale *COCO* dataset, where we obtain significantly better results with as few as 1% of labeled annotations.

Part III

Applications, conclusions and perspectives

7 Potential applications

The greatest opportunity offered by AI is not reducing errors or workloads, or even curing cancer: it is the opportunity to restore the precious and time-honored connection and trust - the human touch - between patients and doctors.
– Eric Topol, Deep Medicine



Our approaches can be applied inside as well as outside the OR, for example, in radiation risk monitoring (see [Loy Rodas 2018]), and medical infant motion analysis (see qualitative video https://cutt.ly/mini-rgbd-qual)

Chapter Summary

In this chapter, we discuss some of the potential applications of our approaches inside the Operating Room (OR). We also demonstrate the generality of our approach outside the OR environment by illustrating an initial work on medical infant pose analysis.

7.1 Radiation exposure estimation

As discussed in section 1.3.1.3, the rise of intraoperative X-ray imaging in hybrid surgeries demands the clinicians to remain close to the imaging device and eventually to the potentially harmful ionizing radiations. Recent works such as [Rodas 2016, Krebs 2021] have developed a radiation simulation system utilizing person-body models to measure the radiations at different body parts. Figure 7.1 shows one such application developed in our lab utilizing human pose estimation to calculate radiation exposure at different parts of the body. This intuitive risk awareness application can be further extended by utilizing our pixel-based instance segmentation approach, as discussed in chapter 6, for a more fine-grained assignment to different body parts in the radiation simulation. It would be more clinically relevant not just because of the fine-grained pixel-level radiation assignment to clinicians, but even more because of the more accurate simulation modeling by considering the detailed OR layout.

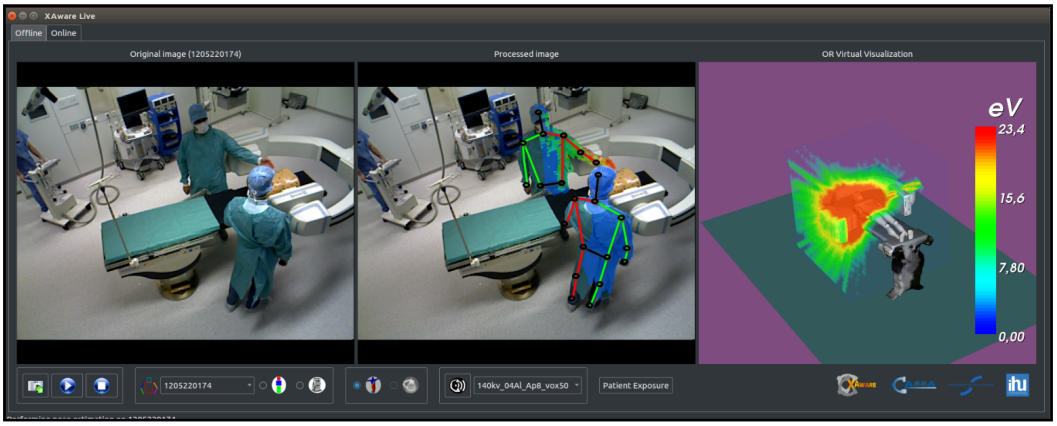


Figure 7.1: Integrating human pose estimation for radiation exposure estimation inside the OR, see [Loy Rodas 2018]

7.2 Surgical skills assessment

As discussed in section 1.3.1.2, it is essential to develop automated ways to evaluate surgeon's skills inside the OR. The current methods of evaluating surgeon skills mainly utilize the tool motion, and surgical workflow analysis [Liu 2021a]. While these provide essential cues for surgical skills, a fine-grained analysis of the surgeon's face and hand motion can provide essential complementary cues. Approaches that can automatically analyze these complementary visual cues could help develop a complete surgical skills evaluation system. Moreover, it could also further help understand the essential non-verbal communications inside the OR to analyze cognitive load, especially during the critical phases of surgery. As an initial work, we develop a Unsupervised Domain Adaptation (UDA) approach to estimate 133 whole-body keypoints with dense 68 face keypoints and 42 hand keypoints. Our approach, similar to ORPose discussed in chapter 5, utilizes the recently introduced COCO whole-body [Jin 2020] as the source domain dataset and the *MVOR-unlabeled* dataset (see section 3.2.1) as a target domain dataset. Figure 7.2 shows some qualitative results for the whole-body pose estimation. The detailed motion analysis of these dense face and hand keypoints can help not only to evaluate essential surgical skills but also to understands non-verbal communications in the OR.

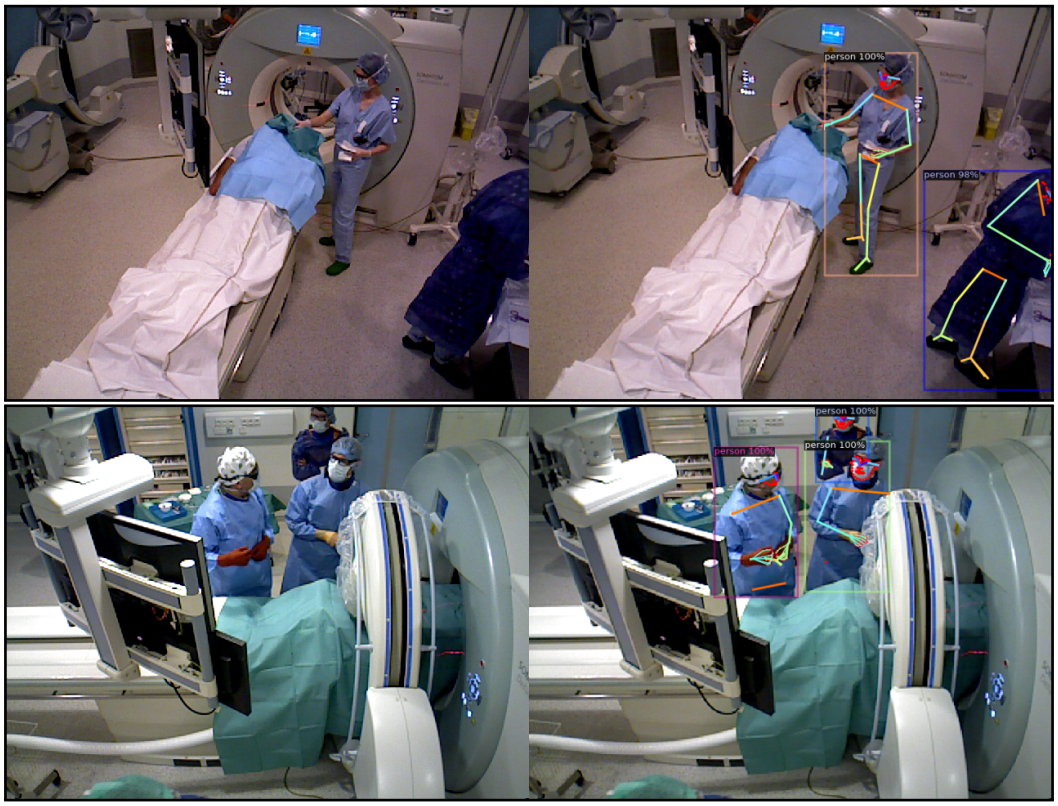


Figure 7.2: Qualitative results for full body pose estimation including dense hand and face keypoints.

7.3 Medical infant motion analysis

Infant motion analysis to assess spontaneous movements at a very young age by trained experts enables early detection of neurodevelopmental disorders like Cerebral Palsy (CP). Authors in [Prechtl 1990] observe that the quality of spontaneous movements of infants is a good marker for detecting impairments of the young nervous system at the age of 2-4 months.

An automated motion analysis system requires accurately capturing body movements, ideally without markers or attached sensors not to affect the movements of infants. A vast majority of recent approaches for human pose estimation focus on adults having very few infants' poses, leading to a degradation of accuracy when applied to infants. Specifically in the clinical setting, it is particularly hindered due to privacy constrains in collecting the infants' motion data. Protection of privacy of infants has been more strict than adults, as infants can not decide whether or not they want their image to be published.

As our initial experiments in this direction, we utilize the first publicly available synthetic dataset, called MINI-RGBD [Hesse 2018a], to develop UDA approach for the young infants. As the dataset is synthetically generated containing limited texture and motion various, we observe that an model trained on the *COCO* dataset, called *kmrcnn++* (see section 6.2.2.1) works well on the original size. Our domain adaption approach, called *Adapt-OR* (see section 6.2.3), on the MINI-RGBD dataset, achieves significantly better results on the privacy-preserving low-resolution images.

MINI-RGBD is a synthetic dataset of infants in motion created using the Skinned Multi-Infant Linear (SMIL) [Hesse 2018b] body model specifically designed for infants. It has 12 sequences each consisting of 1000 RGBD-frames containing 2D/3D poses and segmentation masks. We use video sequence "01", "03", "04", "06", "07", "09", "10", "12" for unsupervised training and "02", "05", "08", "11" for testing. We do not use ground truth annotations to train our model. It uses 25 keypoints, however we use 13 common COCO keypoints for testing. We use our *AdaptOR* approach described in section 6.2.3 to train on the MINI-RGBD dataset. We train the model for 10k iterations with same experimental settings as discussed in chapter 6, section 6.4.3.

Table 7.1 shows the quantitative results. We observe a slight improvement at the original resolution, most likely due to the synthetic nature of the data. However, we obtain significantly better results on the heavily down-sampled low-resolution with

Table 7.1: Quantitative results on the MINI-RGBD dataset for before and after domain adaptation.

Model	$\mathbf{AP^{bb}_{person}}$		$\mathbf{AP^{kp}_{person}}$	
	1x	12x	1x	12x
source-only(kmrcnn++)	87.86	81.85	94.22	51.17
AdaptOR	87.94	85.25	95.59	68.55

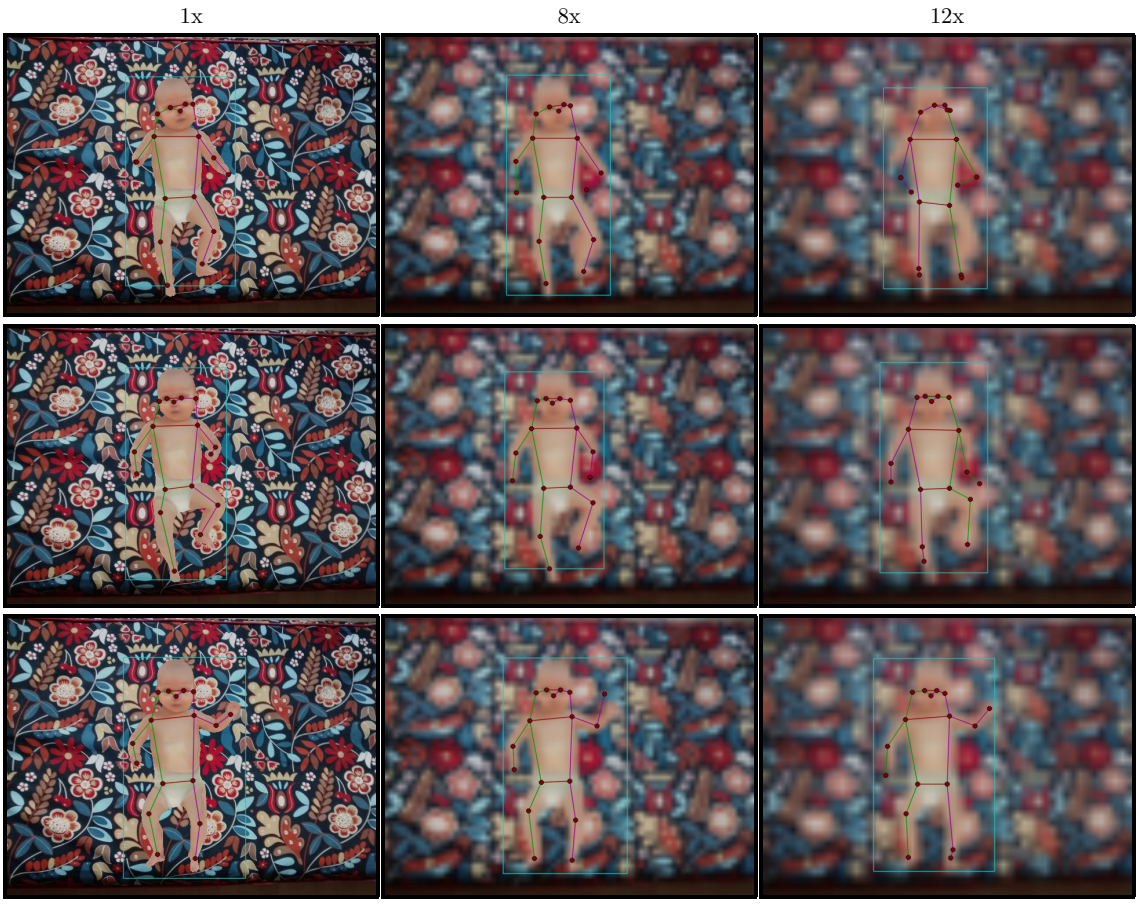


Figure 7.3: Qualitative results on the MINI-RGBD dataset at various low resolution

about 4% improvement in bounding box AP and 17% improvement in keypoint AP. Figure 7.3 show the qualitative results on some sample MINI-RGBD frames at various downsampled scales (1x, 8x, 12x).

8 Conclusions

The next AI revolution will not be supervised.
– Alyosha Efros

Chapter Summary

We finally conclude this dissertation by summarizing the contributions of our work. We also discuss the possible directions of future research to address the current limitations and improve the performance of our methods. We also discuss the new ideas originating from these works which could enable novel applications inside the Operating Room (OR).

8.1 Conclusion

The fine-grained localization of clinicians in the OR is a key component in designing the new generation of OR support systems. Computer vision models for pixel-based segmentation and body-keypoint detection are needed to understand the OR's clinical activities and spatial layout. This task, however, poses challenges in the following two dimensions: 1) significant visual domain differences of OR images w.r.t traditional vision datasets, and 2) privacy barriers in collecting and generating the data and annotations. The principal goal of this dissertation is to develop approaches that exploit large-scale unlabeled OR data for fine-grained person localization while preserving the privacy of clinicians.

As our first contribution, we introduce a new Multi-view Operating Room (MVOR) dataset, the first public dataset recorded during real clinical interventions. The MVOR dataset illustrates the inherent visual challenges at a *global-level* showing significant color distribution variations and at a *instance-level* showing challenging clinicians' poses due to proximity, loose clothes and masks, and multiple occlusions from instrument clutter. The MVOR dataset contains 2196 Red-Green-Blue-Depth (RGBD) frames from 3 synchronized cameras and annotations for multi-person detection and 2D/3D human pose estimation. Given the small size of the MVOR dataset, it is proposed to serve as a *test-bed* to evaluate a model's ability to generalize to unseen configurations and color distribution of the real-world data. The ground-truth annotations of the dataset are subsequently updated to the COCO format for standardized evaluation and compared against the model trained on the natural images. We set up our baselines on the MVOR color and depth images at original size images (1x) and heavily downsampled images (12x) by evaluating several current state-of-the-art methods for person bounding box detection, instance segmentation, and 2D/3D human pose estimation. As expected, the results show a substantial decrease in the accuracy at the original scale (1x) and significantly poor accuracy at the downsampled scale (12x). These initial baseline evaluation results show a large margin of improvement on the challenging OR images, specifically on the privacy-preserving low-resolution images.

In the chapter 4, we present our first Unsupervised Domain Adaptation (UDA) approach for multi-person human pose estimation on low-resolution depth images. We propose a novel method to perform domain adaptation across visual modalities of color (RGB) images to depth (D) images. We exploit the synchronized properties of the RGBD images and the recent pose estimator's performance on the RGB images to propose a *pseudo-labeling* based approach. We propose using the model's prediction on the high-resolution RGB images as pseudo labels for the corresponding low-resolution depth images. Our approach uses the high-resolution color images only during the training thus can be applied in the clinical setting where we need the model for the low-resolution depth images. We evaluate the effectiveness of our approach of generating automatic pseudo labels on the two types of human pose estimation models: on a *bottom-up* RTPose where we integrate super-resolution feature blocks to effectively for

low-resolution feature learning and on a *top-down* Keypoint-RCNN where we propose to exploit advance data-augmentations for the effective low-resolution feature learning. Our evaluation of these two methods on *MVOR* dataset shows that even with a 12x sub-sampling of the depth images, our method achieves results better than the pose estimator trained on original-size RGB images. These results suggest the high potential of low-resolution images for scaling up and deploying privacy-preserving AI assistance in hospital environments. We further show that improved underlying HPE architecture and strong data augmentations significantly boosts the performance and effectively learns the features for low-resolution images.

In chapter 5, we propose a UDA approach for joint 2D/3D pose estimation on monocular RGB OR images. We propose to use a *teacher-student* based *pseudo-labeling* paradigm where a multi-stage, larger, and accurate teacher model generates pseudo labels for a single-stage, smaller, and practical student model. We propose to generate accurate pseudo labels using the teacher model by combining advances from the *data-distillation* and *knowledge-distillation*. The *data-distillation* has been utilized by averaging predictions from the multiple augmentations of the same instance to reduce the pose ambiguities, and *knowledge-distillation* has been utilized by incorporating *hard* as well as *soft* averaged predictions in the training loop. The *hard* predictions contain high confidence detections, whereas *soft* predictions contain all the predictions along with their confidence value. This effective strategy to generate accurate pseudo labels greatly benefits the student model during the training. Our evaluation on the *MVOR+* dataset demonstrates that the proposed network can yield accurate results on low-resolution images, as needed to ensure privacy, even using a downsampling rate of 12x.

As our last contribution, in chapter 6, we tackle the joint person pose estimation and instance segmentation task and propose a UDA approach to adapt a model trained on a labeled source domain to an unlabeled target domain. We propose a new *self-training* based framework with advanced data augmentations to generate pseudo labels for the unlabeled target domain. The high-quality effectiveness in the pseudo labels is ensured by applying the explicit geometric constraints of the different augmentations on the unlabeled input image. We also introduce disentangled feature normalization where we use different feature normalization layers to tackle two statistically different source and the target domains distributions. The training is further stabilized by utilizing *mean-teacher* paradigm where we use closely coupled teacher and student models to generate and consume the pseudo labels, respectively. Evaluation on the two target domain datasets, *MVOR* and *TUM-OR-test*, with extensive ablation studies, shows the effectiveness of our approach. We further demonstrate that the proposed approach can effectively be adapted to the low-resolution images of the target domain, as needed to ensure OR privacy, even up to a downsampling factor of 12x. Finally, we illustrate the generality of our approach as Semi-supervised Learning (SSL) method on the large-scale *COCO* dataset, where we obtain significantly better results with as few as 1% of labeled annotations.

8.2 Perspectives

We believe that several possible lines of research could spawn from work presented in this dissertation exploiting large-scale unlabeled data in novel ways. These could open up possibilities for exploring exciting research problems, consequently enabling the OR of the future as the OR of the present. Several perspectives of our work are discussed below.

8.2.1 Multi-view multi-person absolute 3D pose estimation

In this dissertation, we have introduced the single-view person localization approaches utilizing only the unlabelled data from the OR. However, in a single view, it is possible that the clinicians can hardly become visible due to the heavy occlusion. It could eventually result in poor performance, specifically in the 3D pose estimation, which relies on accurate 2D pose estimation input, see figure 8.1. Multi-view images can not only help to resolve the occlusion failures but also help to provide absolute 3D pose estimation in the OR room coordinates as opposed to root-relative coordinates predominantly used in single-person 3D pose estimation. The current research for

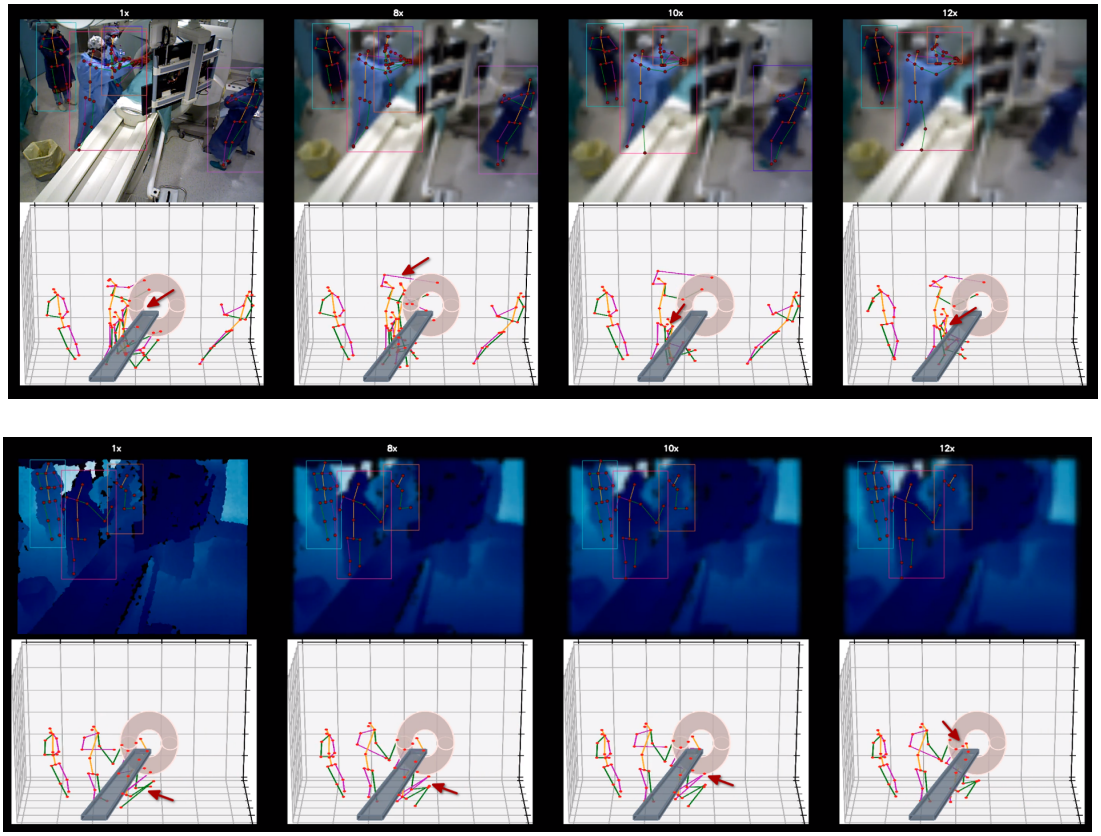


Figure 8.1: Some of the failure cases of our approaches in 3D pose estimation on color and depth images arising mainly due to the heavy occlusion from other clinicians and instrument clutter. Multi-view images can help to resolve these failure cases by taking complementry cues from other views.

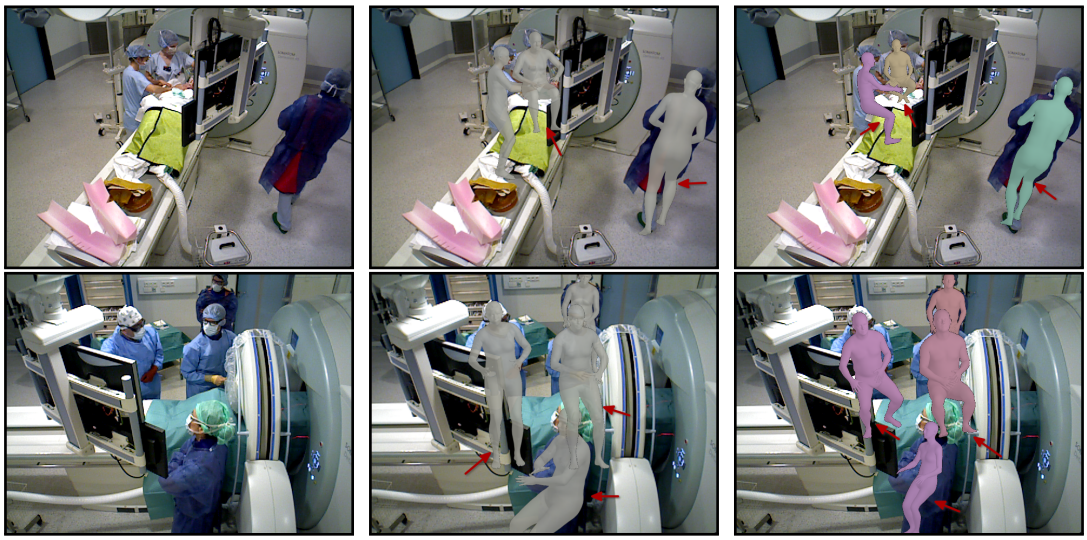
multi-view multi-person absolute 3D pose estimation is however limited to fully supervised approaches [Tu 2020, Reddy 2021] on the datasets recorded during simulated social activities [Joo 2015]. Utilizing UDA framework to extend these fully supervised approaches to an unseen unlabeled real-world OR environment with a limited number of camera views for accurate absolute 3D pose estimation is a challenging future research direction to pursue. More interestingly, another promising direction would be learning absolute 3D pose estimation in a purely self-supervised manner. When multi-view images are available, the 3D pose in absolute coordinates can be determined by using multi-view geometry, and triangulation [Andrew 2001]. However, the bottleneck lies in the inaccuracy of 2D pose estimations due to occlusion and limited camera views. The spatial and geometric constraints as discussed in chapter 6 can be extended in the multi-view volumetric space to iteratively refine the 3D poses by exploiting these constraints, triangulation, and more robust 2D human pose estimation.

8.2.2 Multi-modality person localization using RGBD images

This thesis has primarily focused on the use of unlabeled data to develop various fine-grained person localization approaches. However, the developed approaches take single modality (color or depth) input images at the inference time. With the advent of inexpensive RGBD cameras, the depth and color images are readily available, providing complementary information of the scene - 2D texture details from the color and 3D environment details from the depth images. The Convolutional Neural Network (CNN) based models have shown significantly improved performance for the 2D color images. However, the straightforward application of CNN models on the depth images may not be optimal. An interesting research direction would be to explore a unified architecture for RGBD images exploiting complementary information from both the modalities at the inference time. Such an architecture can help not just to improve accuracy at the inference time but even more to extend UDA approach, such as discussed in chapter 6, for more robust domain adaptation.

8.2.3 Exploiting temporality for accurate pose estimation and consistent tracking

Temporality is another dimension that can be exploited to improve further and stabilize the localization accuracy. For example, the *teacher-student* based UDA framework described in chapter 5 can exploit temporality in the teacher model to get more accurate pseudo labels. The student model can further be extended with the temporal component to exploit these accurate pseudo labels to get more accurate output and consistently track the clinicians. As another example, the UDA framework described in chapter 6 can also exploit the temporality to more accurately generate and consume the pseudo labels for the temporal sequence.



Input CenterHMR [Sun 2020] VIBE [Kocabas 2020]

Figure 8.2: Failure cases of some of the state-of-the-art approaches for human body pose and shape estimation when applied to the challenging OR images.

8.2.4 Human body pose and shape estimation

This thesis has primarily worked on the fine-grained person localization ranging from person instance segmentation to 2D and 3D human pose estimation. Although extremely useful, the instance segmentation is limited to the 2D image, and pose estimation localizes sparse keypoints either in 2D or 3D. The current research directions have been evolving to estimate parametric 3D human mesh model from a given image combining the pixel-based instance segmentation and the 3D pose estimation into a unified field [Loper 2015, Kanazawa 2018]. The research however has primarily been focused on learning the body shape model from synthetic datasets such as Human3.6 [Ionescu 2013]. When applied to challenging real-world environments such as OR, these approaches fail remarkably, as illustrated in figure 8.2. The UDA frameworks discussed in this dissertation could help enable bringing human body shape estimation inside the OR.

List of Publications

International journals

Vinkle Srivastav, Afshin Gangi and Nicolas Padoy, *Unsupervised domain adaptation for clinician pose estimation and instance segmentation in the OR.*, Submitted to Medical Image Analysis, 2021.

Qualitative results video: https://youtu.be/gqwPu9-nfGs

Project page: https://github.com/CAMMA-public/hpe-adaptor

ArXiv link: https://arxiv.org/abs/2108.11801

Thibaut Issenhuth, Vinkle Srivastav, Afshin Gangi, and Nicolas Padoy, *Face detection in the operating room: Comparison of state-of-the-art methods and a self-supervised approach.*, International journal of computer assisted radiology and surgery 14, no. 6 (2019): 1049-1058.

International conferences with proceedings

Vinkle Srivastav, Thibaut Issenhuth, Abdolrahim Kadkhodamohammadi, Michel de Mathelin, Afshin Gangi, and Nicolas Padoy, *MVOR: A multi-view RGB-D operating room dataset for 2D and 3D human pose estimation*, MICCAI workshop on Large-scale annotation of Biomedical data and expert label synthesis (MICCAI-LABELS), 2018

Project page: https://github.com/CAMMA-public/mvor

Vinkle Srivastav, Afshin Gangi and Nicolas Padoy, *Human pose estimation on privacy-preserving low-resolution depth images*, Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2019

Qualitative results video: https://cutt.ly/depthpose

Project page: https://github.com/CAMMA-public/orpose-depth

Vinkle Srivastav, Afshin Gangi and Nicolas Padoy, *Self-supervision on Unlabelled OR Data for Multi-person 2D/3D Human Pose Estimation*, Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2020

Qualitative results video: https://cutt.ly/orpose3d

Project page: https://github.com/CAMMA-public/orpose-color

Book chapter

Deepak Alapatt and Pietro Mascagni, Vinkle Srivastav, and Nicolas Padoy *Neural Networks and Deep Learning.* In Hashimoto D.A. (Ed.) Artificial Intelligence in Surgery. New York: McGraw Hill. ISBN: 978-1260452730
Project page: https://github.com/CAMMA-public/AI4surgery

Appendices Part IV

A Face Detection in the Operating Room: Comparison of State-of-the-art Methods and a Self-supervised Approach

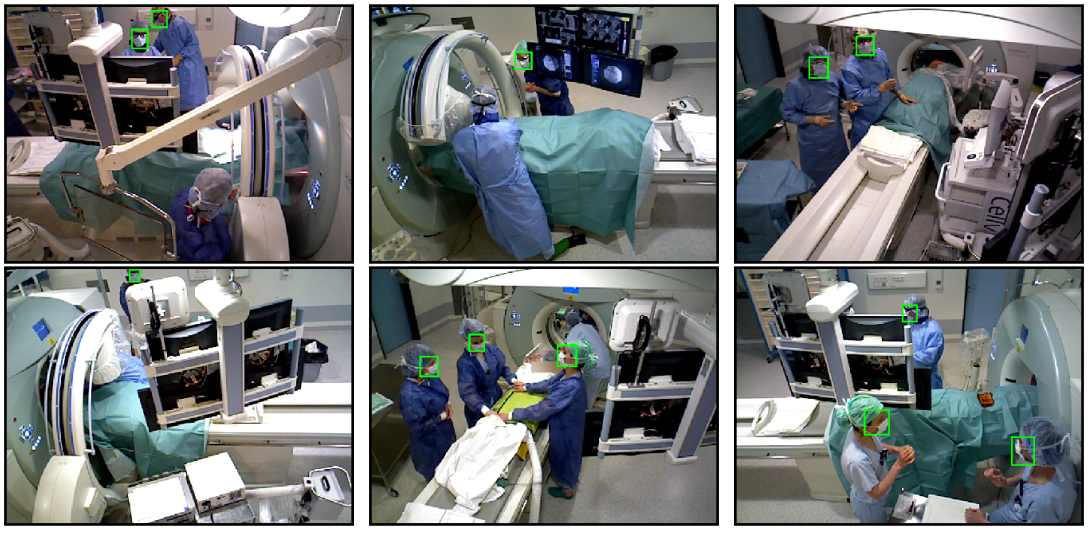


Figure A.1: Examples images from the MVOR-Faces dataset collected in the OR, illustrating the challenges for face detection systems: occlusion with medical equipment or other persons, masked faces, absence of visible skin. Ground-truth is shown with green bounding boxes. Our proposed self-supervised approach significantly improves face-detection results on the challenging OR images.

Appendix A. Face Detection in the Operating Room: Comparison of State-of-the-art Methods and a Self-supervised Approach

Chapter Summary

In this chapter, we describe our work on face detection in the OR. This work was carried out by Thibaut Issenhuth (currently the Ph.D. candidate at Criteo and Ponts ParisTech) during an internship, which I co-supervised with Prof. Nicolas Padoy. It was presented at the international conference on information processing in computer-assisted interventions (IPCAI) conference and published in the international journal of computer-assisted radiology and surgery (IJCARS) [Issenhuth 2019].

A.1 Introduction

Face detection in the operating room is one of the key steps needed to develop intelligent context-aware systems for the automatic analysis of human activities. It can indeed serve for person detection and identification as well as for the anonymization of sensitive OR data.

Face detection is a very active research topic in computer vision. Before the rise of deep learning, traditional methods for face detection used machine learning algorithms on top of hand-crafted features [Viola 2001]. With the advent of deep learning architectures based on convolutional neural networks (CNNs), the performance of face detectors has drastically improved. CNNs are trained end-to-end and are able to learn semantically rich and robust data representations that yield great accuracy. The face detection architectures are often inspired by deep object detectors, whether they are one-stage [Najibi 2017, Zhang 2017] or two-stage detectors [Jiang 2017]. The one-stage detectors generally divide the image into a grid of boxes and directly classify and regress the localization of objects in each box. The two-stage networks first use a Region

Proposal Network (RPN) [Ren 2015] to extract Region of Interests (ROIs), then a second network to classify and localize each ROI more accurately. These detectors manage to handle the variety of scales, by setting up strategies to specifically detect small faces. They perform contextual reasoning and use image or features pyramids to achieve robustness. The success of these methods can also be attributed to the availability of large-scale annotated dataset. Indeed, WIDER Faces [Yang 2016], the standard dataset for training and testing face detection methods in the wild, contains 32,203 images with 393,703 labeled faces. Apart from the bounding box based face detection methods, faces can also be extracted from the face keypoints of human pose estimators, which is performed by mainly two types of approaches: bottom-up and top-down. Bottom-up approaches [Cao 2017, Insafutdinov 2016] first detect all the keypoints, then assemble them into skeletons, whereas top-down approaches [Fang 2017, Xiao 2018, Chen 2018a] first detect persons, often with standard object detectors, and then detect keypoints for each detected person using a single person pose estimator. The top-down approaches resolve better the keypoint to person assignment and therefore largely outperform bottom-up approaches on the standard public datasets [Lin 2014, Andriluka 2014].

The automatic recognition of activities during real surgeries to develop intelligent context-aware assistance systems is a recent field that has started to gain traction in the medical as well as computer vision community [Twinanda 2016b, Maier-Hein 2017, Yeung 2018]. Work on analyzing humans in OR videos have generally focused on person bounding box detection and on human pose estimation, using either RGB or RGB-D data [Kadkhodamohammadi 2017b, Kadkhodamohammadi 2017c, Belagiannis 2016]. So far, face detection in the OR has however received very little attention, besides [Nieto-Rodríguez 2015] and [Flouty 2018], described below. Furthermore, current state-of-the-art face detectors, even those close to the human-level performance, do not generalize well to OR images. Their inferior generalization can be explained by the fact that they have been trained on natural images, whereas OR images are very specific and challenging: persons' faces are often occluded due to equipment clutter, masks, and glasses. Figure A.1 shows some examples of the challenging situations occurring inside the OR.

One standard approach to overcome this visual domain difference is to use an annotated dataset from the target application to either retrain a fully supervised method or adapt an existing method by fine-tuning its parameters through transfer learning. For example, in [Nieto-Rodríguez 2015] models based on AdaBoost [Friedman 2000] are trained from scratch to detect faces and the absence/presence of masks. Whereas in [Flouty 2018], authors recently proposed a method to detect the faces in the OR by finetuning the Faster-RCNN detector [Jiang 2017] on OR videos. In [Flouty 2018], the video dataset consists of youtube OR videos, which have been manually annotated with face bounding boxes. The results are further improved by using temporal smoothing. Manual annotations, as required by the aforementioned approaches, can however be expensive and time-consuming, whereas non-annotated data is in abundance and often inexpensive.

Therefore, this work aims at distilling knowledge from non-annotated OR data to improve the baseline performance of a face detector.

This chapter investigates face detection and visual domain adaptation for the OR environment. We first present a comparison of 6 state-of-the-art face detectors. We consider methods where faces can be obtained either directly from bounding box based methods or from face keypoints generated by human pose estimators. We evaluated these methods on the MVOR-Faces, an extension of the MVOR dataset [Srivastav 2018] augmented with face bounding box annotations. To the best of our knowledge, this chapter presents the first comparison of state-of-the-art face detectors in an OR environment. We also select one detector, SSH [Najibi 2017], and propose to improve it by using an iterative self-supervised method. Several variants of self-supervised methods have been recently used to improve the quality of the synthetic annotations, for instance by using temporal ensembles [Laine 2016] or by combining the results of different geometric transformations [Radosavovic 2018]. In this work, we found it effective to iteratively generate synthetic annotations and fine-tune the model. This approach significantly improves the original model, and largely outperforms the best face detectors on all metrics.

A.2 Methodology

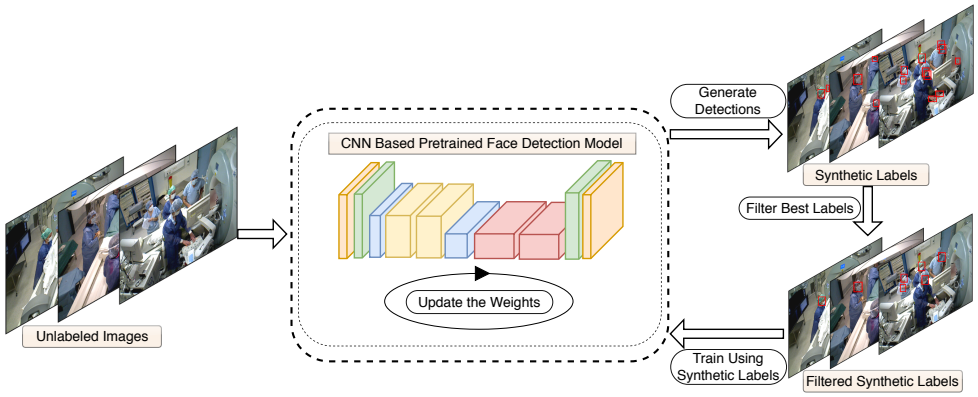


Figure A.2: An iterative approach to adapt a face detector to a target operating room. First, we obtain a trained face detector (SSH [Najibi 2017]) and unlabeled images of the same operating room. Then, we repeat the following steps: (a) use the detector to generate the labels (b) filter the detections with a heuristic to create good quality synthetic annotations (c) retrain the detector using synthetically generated annotations.

A.2.1 Comparison of state-of-the-art face detector

We present below the state-of-the-art methods for face detection used in our comparison.

In this study, the faces are represented by bounding boxes. We consider 4 methods where the faces are directly obtained as the output of the detector and 2 methods where the face bounding boxes are generated from face keypoints detected by human pose estimators. These methods are selected based on their ranking on standard public

datasets, namely the WIDER dataset [Yang 2016] for bounding box based face detectors and the COCO dataset [Lin 2014] for human pose estimators. For reproducibility, we only choose open-source methods.

A.2.1.1 Bounding box based face detectors

Faster-RCNN face detector [Jiang 2017]. The Faster-RCNN, originally designed as a generic two-stage object detector, was trained for the face detection task on the WIDER Faces dataset. First, the RPN generates ROIs with a sliding window approach on deep feature maps. At each sliding window location, anchors, bounding boxes of different scales and aspect ratios, are predicted as either background or ROI. Then, ROIs are pooled and used as input for a second network, which classifies the face and regresses for the exact coordinates of its bounding box.

Finding tiny faces [Hu 2017]. This method is specifically conceived to detect faces of different scales. The input of the algorithm is an image pyramid, with three versions of the image: one downsampled, the original and one upsampled image. Each rescaled image is processed by a shared pyramidal CNN, which predicts binary heatmaps for bounding box templates of different sizes.

SSH: Single stage headless face detector [Najibi 2017]. This is a one-stage face detector that includes a context module, namely a set of convolutional layers to increase the effective receptive field and different branches to achieve scale-invariance. It uses three different detector networks to predict small, medium and large face anchors. SSH achieves a similar accuracy than the tiny face detector [Hu 2017], while maintaining real-time performance.

S^3FD: Single shot scale-invariant face detector [Zhang 2017]: This is also a one-stage face detector inspired by the RPN [Ren 2015] and SSD [Liu 2016] architectures. They design strategies to increase the number of positive anchors matching tiny faces during training. Their CNN architecture includes feature maps and detectors that are specific to a range of scales and uses a max-out background label on small anchors to reduce the number of false positives.

A.2.1.2 Human pose estimation based face detectors

Human pose estimation aims to localize the anatomical keypoints of all the persons present in an image. These anatomical keypoints are spread across the whole body including the face keypoints. Current state-of-the-art methods are trained on the COCO dataset, which includes five face keypoints (nose, left eye, right eye, left ear, right ear). We fit a bounding box of fixed size, 30x30 pixels, at the average location of these five keypoints to extract the face bounding box. We now describe the pose estimation methods that we used for the evaluation.

OpenPose [Cao 2017]. This is one of the best bottom-up approaches for human pose estimation. First, a CNN predicts confidence heat-maps for all keypoints and part affinity fields for each joint. These maps are then used to construct a graph of keypoints

and body joints. Finally, this graph is parsed using a bi-partite graph matching algorithm to produce a set of human poses.

AlphaPose [Fang 2017]. This is one of the state-of-the-art top-down methods. It uses Faster-RCNN [Ren 2015] to detect persons. Then, cropped human bounding boxes are processed by a single person pose estimator, which is composed of several modules, including spatial transformers, to refine the keypoint detections in the bounding box. This method successfully handles the problems caused by inaccurate and duplicate bounding boxes.

A.2.2 Iterative self-supervised approach for face detection in the OR

We use the following two steps to improve the selected state-of-the-art face detector on OR images: 1. Generation of the unlabeled dataset 2. Iterative refinement using a self-supervised approach.

A.2.2.1 Generation of the unlabeled dataset

We use an unlabeled dataset of 20k images generated from videos captured in the OR. The videos were collected on days different from the ones of the test dataset to ensure the absence of overlap between the unlabeled dataset and the test dataset. We then use OpenPose [Cao 2017], a multi-person pose estimator, on the OR videos to get the approximate number of persons in each frame. The computational efficiency of OpenPose allows us to make the inference on the entire dataset in a reasonable time. We divide the images into four categories: images with one, two, three, and four or more detected persons. Since OpenPose also gives a confidence score for each detected skeleton, we average the scores of the detected skeletons and take the 5k highest-scored images from each category (i.e., 20k images overall). This selection method ensures that the images contain persons in different numbers.

A.2.2.2 Iterative refinement using self-supervised approach

We utilize an iterative self-supervised approach to adapt the state-of-the-art model to the target OR dataset. This approach consists of fine-tuning the model on a subset of its own detections. We use SSH [Najibi 2017], pre-trained on WIDER Faces, as the CNN-based model for the iterative refinement. We choose this model because it has high computational efficiency and also yields state-of-the-art results on WIDER Faces. This detector is then used to generate synthetic labels on the unlabeled dataset. To select quality face bounding boxes, we use a simple yet effective heuristic criteria: with a dataset of N images, we select the best 2*N detections. Since we have approximately 2.5 persons/image in the unlabeled dataset, 2*N best detections contain the face bounding boxes with a high recall. These synthetically annotated images are then used to finetune the original model. We perform these steps iteratively to improve the detections and the detector at each iteration as shown in Fig. A.2. It is to be noted that no validation set is available as we did not use any supervised annotation. Therefore, our experiments

differ from the traditional deep learning experiments, which fine-tune the hyper-parameters based on the performance on a validation set. We mainly conduct the fine-tuning experiments with a different number of training batches before relabelling and different iteration numbers. We present the result of each experiment on the test-set.

A.3 Experimental setup

A.3.1 Test dataset (MVOR-Faces)

We compare the state-of-the-art face detectors on MVOR-Faces, a dataset of operating room images captured during real surgical procedures. MVOR-Faces is an extension of the public MVOR dataset [Srivastav 2018], which consists of 732 multi-view frames (2196 images) recorded in an interventional room. In the MVOR dataset, faces of persons without a mask and nude parts of patients are fully blurred, and the persons with masks are blurred only on the eyes. MVOR-Faces contains the same images as MVOR, except that the eyes of the persons wearing a mask are not blurred. Also, it contains the manually annotated face bounding box of all visible faces wearing a mask. All fully-visible faces and nudity zones are still blurred in the MVOR-Faces as needed for anonymity. Overall, the dataset contains 2262 face bounding boxes for 2196 images.

A.3.2 Evaluation metrics

We use the standard metrics for object detection from COCO [Lin 2014], i.e. Average Precision (AP) and Average Recall (AR). AP^{IoU} is the average precision at a fixed intersection over union (IoU), and AP is the average of AP^{IoU} at different IoU thresholds. While the public implementation averages between IoU of 0.5 and 0.95 with a step of 0.05, we average between an IoU of 0.3 and 0.95 to support a slightly looser metric. The consideration of a slightly looser metric is motivated by the fact that face detection in a medical context is quite challenging: clinicians wear mask, glasses, and hats, and are often occluded. Therefore, a looser metric reduces the bias in favor of the face detectors.

A.4 Results and discussion

A.4.1 Comparison of state-of-the-art face detectors

Table A.1 shows the comparison of state-of-the-art face detectors. The first four methods in Table A.1 directly output face bounding boxes, as described in section 2.1.1. The next two methods detect the human skeletons, including face keypoints (i.e. ears, eyes and nose). We extract the face bounding boxes from face keypoints as specified in Section 2.1.2. Unless otherwise stated, we use the exact same models provided by the authors, without modifying any hyper-parameter.

On AP(0.3:0.95), Tiny Face detector [Hu 2017] is the best model, with 0.340; SSH [Najibi 2017] and S3FD [Zhang 2017] are close, with respectively 0.314 and 0.302.

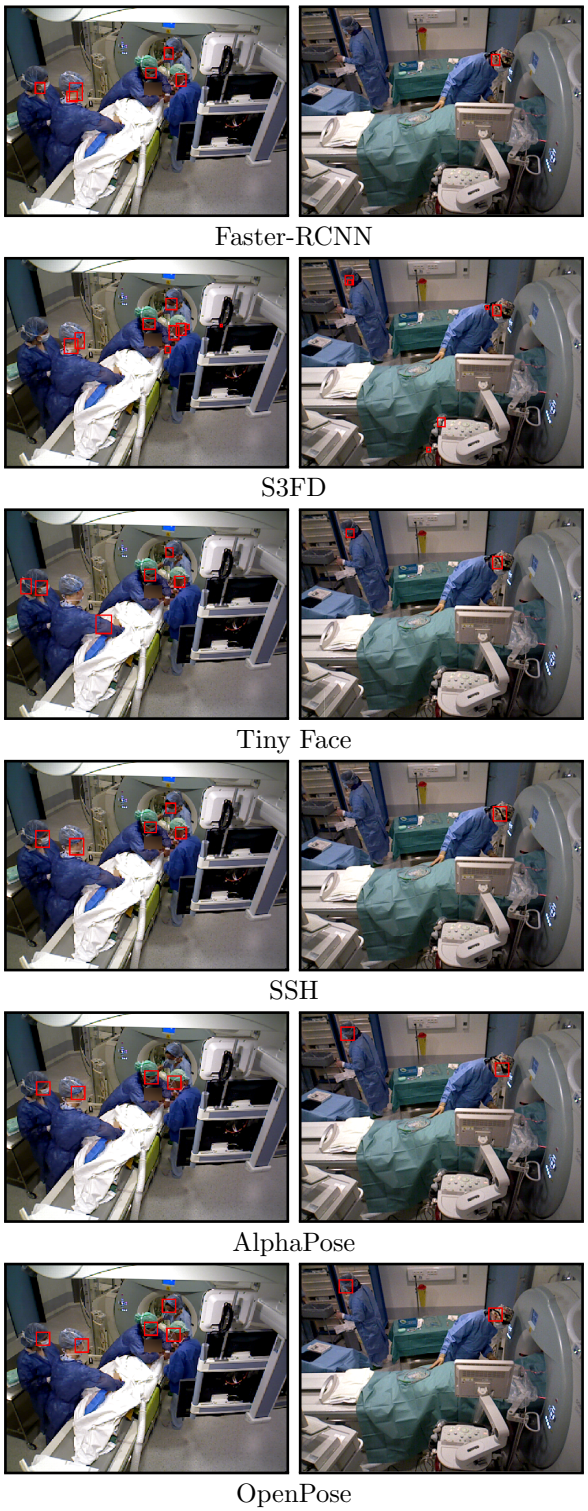


Figure A.3: Qualitative results from the face detectors evaluated on MVOR-Faces. The displayed detections were selected based on a score threshold of the detector corresponding to a recall threshold of 70% at an IoU of 0.3.

Figure A.4: Comparison of the original SSH model (left column) with the best self-supervised model trained with our iterative approach (right column). To filter the displayed detections, we use the score threshold corresponding to a recall threshold of 70% at an IoU of 0.3, as in Fig. A.3. The self-supervised model detects much harder examples, with occlusion or uncommon poses.

On AP(0.3), AlphaPose is the best model with 0.785. With this looser metric, human pose estimators perform better. Indeed, in the OR environment, when clinicians wear

Appendix A. Face Detection in the Operating Room: Comparison of State-of-the-art Methods and a Self-supervised Approach

Detector	AP(0.3:0.95)	AP(0.3)	AP(0.5)	AR(0.3:0.95)
Faster-RCNN [Ren 2015, Jiang 2017]	0.254	0.651	0.407	0.345
S3FD [Zhang 2017]	0.302	0.627	0.486	0.395
Tiny Face [Hu 2017]	0.340	0.734	0.556	0.428
SSH [Najibi 2017]	0.314	0.704	0.517	0.421
AlphaPose [Fang 2017]	0.279	0.785	0.463	0.358
OpenPose [Cao 2017]	0.240	0.776	0.365	0.316
Self-supervised SSH	**0.402**	**0.800**	**0.648**	**0.474**

Table A.1: Results of state-of-the-art face detectors on MVOR-Faces. First four methods are bounding box based face detectors. AlphaPose and OpenPose are human pose estimators, from which face bounding boxes are generated from the face keypoints. Results show the margin for improvement on the MVOR-Faces dataset.

Detector	AP(0.3:0.95)	AP(0.3)	AP(0.5)	AR(0.3:0.95)
Tiny Face [Hu 2017]	0.237	0.627	0.369	0.331
SSH [Najibi 2017]	0.229	0.600	0.368	0.368
AlphaPose [Fang 2017]	0.239	**0.742**	0.370	0.323
Self-supervised SSH	**0.306**	0.711	**0.492**	**0.403**

Table A.2: Comparative study: results of state-of-the-art face detectors on *MVOR* [Srivastav 2018], the public version of MVOR-Faces. Here, clinicians wearing a mask are blurred around the eyes. Results show the significant decrease in the performance as compared to MVOR-Faces shown in Table A.1

mask and hats, face detectors cannot rely on the same features as in the outside environment, such as the mouth shape and the nose. Human pose estimators, which also detect other body keypoints, are more robust than face detectors. However, they do not localize the bounding boxes accurately enough to perform well on stricter metrics. For

Figure A.5: Failure examples from the self-supervised model showing best results on the test-set, i.e. MVOR-Faces. False positives mainly come from hands, which are mistaken as faces. False negatives arise with hard examples, i.e. faces strongly occluded or with difficult poses.

	AP(0.3:0.95)	AP(0.3)	AP(0.5)	AR(0.3:0.95)
Original SSH model results → Number of training batches ↓	0.314	0.704	0.517	0.421
1000	0.372	**0.781**	0.597	**0.462**
2000	0.373	0.769	0.593	0.458
3000	0.374	0.770	0.595	0.458
5000	0.372	0.770	0.598	0.454
10000	**0.378**	0.778	**0.608**	0.462
15000	0.372	0.781	0.600	0.457

Table A.3: Comparative study: training SSH model without re-generating the synthetic labels. One training batch is composed of two images. Here, the self-supervised model is trained on the images annotated by the original SSH model and initialized with SSH weights. The AP saturates quite fast. After 1k batches, the AP does not significantly increase. The best results are much lower than best results with the iterative approach in Figure A.2.

Iteration		AP(0.3:0.95)	AP(0.3)	AP(0.5)	AR(0.3:0.95)
	Original SSH model Results → Number of training batches before relabelling ↓	0.314	0.704	0.517	0.421
1	1000	0.373	0.782	0.598	0.462
2		0.365	0.783	0.590	0.461
3		0.367	0.785	0.592	0.450
4		0.365	0.788	0.594	0.456
1	2000	0.373	0.773	0.596	0.459
2		0.385	0.796	0.622	0.466
3		0.393	**0.808**	0.630	0.465
4		**0.402**	0.800	**0.648**	**0.474**
1	3000	0.373	0.772	0.595	0.457
2		0.377	0.793	0.608	0.452
3		0.383	0.806	0.614	0.459
4		0.371	0.797	0.604	0.448

Table A.4: The iterative process of self-supervision with different hyper-parameters. One iteration consists of three steps: (1) Generate predictions on the unlabeled dataset with the last model. (2) Filter detections: select the best 2N detections on N images. (3) Retrain the model on 1k, 2k or 3k training batches. When training with 2k or 3k batches before relabelling, it improves the results from the baseline approach (see in Table A.3). One batch is composed of two images.

comparison, we also provide the results on the original MVOR dataset, where the eyes are blurred, in Table A.2. The results show a significant drop in the performance highlighting the importance of the eyes for face detection.

Overall, results of state-of-the-art detectors show a large margin for improvement on the MVOR-Faces dataset. With an IoU of 0.5, which is a less strict metric, the AP of the best model is only 0.556. On the WIDER Faces dataset, tiny face detector [Hu 2017] achieves 0.819 using the same metric, while SSH [Najibi 2017] reaches 0.944 and S3FD [Zhang 2017] 0.958. Qualitative results shown in Fig. A.3 illustrate some of the mistakes made by the state-of-the-art detectors on the MVOR-Faces dataset, e.g. multiple detections, false positives, false negatives.

A.4.2 Iterative self-supervision

As mentioned in section 2.2, we use SSH [Najibi 2017] for the self-supervised process. We conduct several experiments on self-supervision to demonstrate the interest of this iterative approach. During training, we use the following hyper-parameters: stochastic gradient descent with a learning rate of 0.04, momentum of 0.9 and weight decay of $5e^{-4}$. The batch size is 2. Anchors, which correspond to a location (x,y) in the image and a predefined bounding box size (width, height), are considered as positives if their IoU with a ground-truth bounding box is greater than 0.5, as negatives otherwise. During inference, we use an image pyramid of four levels, as the authors. The aspect ratio of each rescaled image is preserved. The weights are initialized with the ones provided by the authors, after training on the WIDER Faces dataset.

In Table A.4, we provide the test-results of our proposed iterative process, with different hyper-parameters (number of iterations and number of training batches used before relabelling). When relabelling, we filter the detections with the same criteria as explained in section 2.2, i.e. 2*N best detections where N is the number of images. The training is done with the same parameters as mentioned above. The model which performs best on the test dataset is achieved at iteration 4 when training with 2000 batches before regenerating the labels. The model outperforms the state-of-the-art with a large margin on all metrics. On AP(0.5), it outperforms tiny face detector [Hu 2017] by more than 9%, and the original SSH model by 13.1%.

In Fig. A.4, we compare a few detections from the original SSH and the best self-supervised model on the test-set. The latter detects much harder examples, with occlusion or uncommon poses, and has fewer false positives. Figure A.5 shows some common failure cases of the self-supervised model. It consists mainly of detecting hands as faces. It shows an overfitting on skin texture. Since the persons wear gowns and masks, faces and hands are almost the only body parts with visible skin. In Table A.3, we show an ablation study of self-supervision for domain adaptation, with no relabelling of target images by the self-supervised model. The 20k images of the unlabeled dataset are annotated with the detections of the original SSH model. We filter the predictions with the same criteria: since we have 20k images, we take the best 40k detections. Then, the model is fine-tuned by training on synthetically annotated images on 15k training batches. We observe a quick saturation on the test-set, MVOR-Faces: the AP(0.3:0.95) reaches 0.372 after 1k batches and 0.378 after 10k batches (i.e., with one epoch on the entire unlabeled dataset). At 15k batches, the AP(0.3:0.95) is back at 0.372. The quick saturation of this process highlights the interest of our iterative approach.

A.5 Conclusion

In this chapter, we propose the first broad evaluation of state-of-the-art face detectors on OR images. Since the results show a large margin for improvement, we also propose to use an iterative self-supervised approach to adapt a face detector to a given OR. It consists of gathering images of the target environment, generating synthetic annotations

with a model trained on a manually annotated dataset, and retraining it iteratively using the synthetic labels. This method is generic and applicable to any OR configuration. Our self-supervised detector outperforms the state-of-the-art on MVOR-Faces by a large margin, namely by more than 6% on AP(0.3:0.95). By significantly improving the accuracy of face detection, we show that self-supervision is a promising direction to transfer state-of-the-art computer vision approaches to the medical context, where annotations are challenging to generate.

B Résumé en français

Approches d'adaptation de domaine non supervisées pour la localisation de personnes dans la salle d'opération

Chapter Summary

B.1 Abstract

L'émergence récente de la science des données chirurgicales promet une nouvelle génération de systèmes d'assistance en salle d'opération (OR, pour "Operating Room"). La localisation précise des cliniciens dans la salle d'opération, soit en termes de points clés en utilisant l'estimation de la pose humaine, soit en termes de pixels en utilisant la segmentation par instances, est un élément clé pour concevoir de tels systèmes. La tâche est cependant difficile non seulement parce que les images OR contiennent des différences visuelles significatives par rapport aux ensembles de données ordinaires en vision par ordinateur, mais aussi parce que les données et les annotations sont difficiles à collecter et à générer dans la salle d'opération, en raison de problèmes de confidentialité. Les approches qui peuvent adapter un modèle à un nouveau domaine cible non étiqueté sont donc très prometteuses.

Dans cette thèse, nous explorons les méthodes d'adaptation de domaine non supervisées pour permettre l'apprentissage visuel pour le domaine cible, la salle d'opération, en travaillant dans deux directions complémentaires. Tout d'abord, nous étudions comment des images basse résolution avec un facteur de sous-échantillonnage allant jusqu'à 12x peuvent être utilisées pour une localisation précise des cliniciens afin de résoudre les problèmes de confidentialité. Deuxièmement, nous proposons plusieurs méthodes auto-supervisées pour transférer les informations apprises d'un domaine source étiqueté vers un domaine cible non étiqueté pour traiter le changement de domaine visuel et le manque d'annotations. Ces méthodes utilisent des prédictions auto-supervisées pour permettre au modèle d'apprendre et de s'adapter au domaine cible non étiqueté. Nous proposons d'abord d'effectuer une adaptation de domaine à travers les modalités visuelles, des images couleur (RGB) vers les images de profondeur (D) en exploitant les propriétés synchronisées des images RGB-D et en utilisant des modèles RGB d'estimation de pose humaine de pointe pour l'estimation sur des images de profondeur. Deuxièmement, nous explorons la distillation des données et celle des connaissances pour générer au mieux des pseudo-étiquettes à partir d'un modèle maître à plusieurs étages, lourd mais précise, afin d'entraîner un modèle élève à un étage, plus petit et déployable pour une estimer la pose humaine 2D/3D. Enfin, nous proposons d'utiliser des contraintes spatiales et géométriques sur les différentes augmentations des couches de normalisation des caractéristiques d'image et de découplage dans le modèle de base pour apprendre simultanément de la source étiquetée et des données du domaine cible non étiquetées pour conjointement estimer la pose et segmenter par instances. Pour démontrer l'efficacité de nos approches, nous publions le premier ensemble de données public, appelé Multi-View Operating Room (*MVOR*), généré à partir d'enregistrements d'interventions cliniques réelles. Nous obtenons des résultats de pointe sur *MVOR*, en particulier sur les images OR à basse résolution préservant la confidentialité. Nous espérons que nos approches d'adaptation de domaine non supervisées proposées pourront aider à développer et à déployer de nouvelles applications d'assistance par IA pour les salles d'opération.

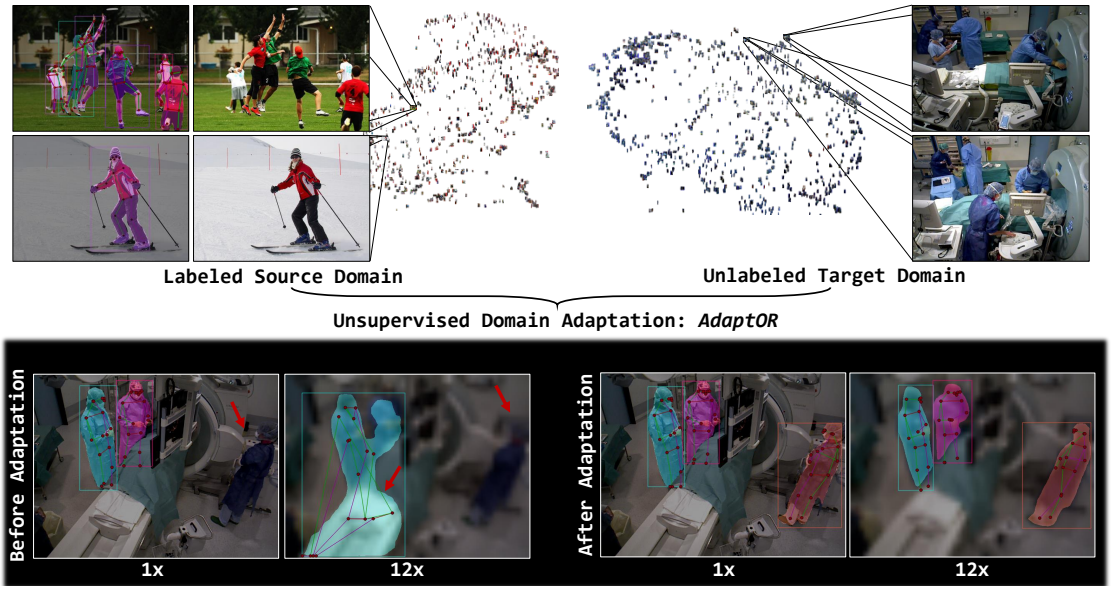
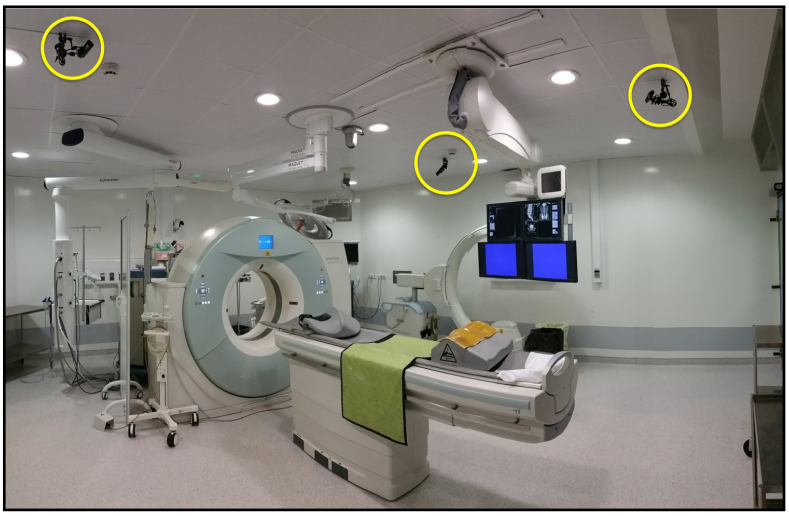


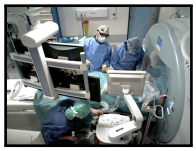
Figure B.1: Différences visuelles globales ainsi qu'au niveau de l'instance entre les images naturelles du *domaine source* et les images OR du *domaine cible*. Lorsqu'un modèle entraîné sur le domaine source est appliqué au nouveau domaine cible, nous constatons une diminution substantielle de la précision de la localisation et une augmentation des détections manquées. Notre méthode d'adaptation de domaine non supervisée améliore considérablement les résultats sur les images OR haute et basse résolution. Les clusters séparés du domaine source et des images du domaine cible sont obtenus en exécutant une technique de réduction de dimension: Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) [McInnes 2018]. Les images des domaines source et cible font partie des ensembles de données COCO [Lin 2014] et MVOR [Srivastav 2018], respectivement.

B.2 Motivation

Les approches de localisation de personne à haut niveau de granularité, telles que l'estimation de la pose humaine à partir de points clés (HPE) ou la segmentation d'instance sur les pixels peuvent fournir dans la salle d'opération (OR) [Kadkhodamohammadi 2017c, Belagiannis 2016] des informations sémantiques de haut niveau sur les cliniciens et les patients. Les nouveaux systèmes d'assistance contextuelle pour la salle d'opération moderne peuvent utiliser ces modèles pour faciliter d'assistance applications telles que la surveillance des activités, la compréhension de la dynamique d'équipe et la surveillance des risques radiologiques. La surveillance de l'exposition aux rayonnements [Rodas 2017], par exemple, pourrait utiliser de tels modèles pour comprendre quel pixel ou point clé des cliniciens est exposé à des rayonnements nocifs. Bien que l'émergence des techniques d'apprentissage profond ait conduit à des modèles très précis pour les images naturelles *in the wild*, la localisation des personnes dans la salle d'opération pose toujours des scénarios difficiles. Les données OR ont une complexité visuelle élevée en raison de l'occlusion et de l'encombrement de divers équipements, des vêtements amples portés par les cliniciens, de la proximité des personnes présentes sur la scène et du changement de distribution des données par

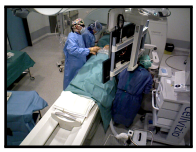


Arrangements de caméras RVB-D au plafond à l'intérieur de la salle d'opération



View 1

View 2



View 3

Figure B.2: Configuration multi-vues et vues acquises dans une salle d'opération de radiologie interventionnelle du CHU de Strasbourg

rapport aux images naturelles. Les méthodes d'apprentissage profond actuelles, qui fonctionnent remarquablement bien pour les images naturelles, montrent une précision réduite lorsqu'elles sont appliquées aux images OR en raison des différences de domaine visuel globales et au niveau de l'instance, comme le montre la Figure B.1. Les ensembles de données supervisés à grande échelle servent de cheval de bataille pour les modèles d'apprentissage profond hautement précis et déployables. Cependant, dans le cas de la salle d'opération, l'obtention d'une annotation de vérité terrain n'est pas triviale, et l'utilisation de techniques de crowdsourcing telles qu'Amazon Mechanical Turk est peu pratique en raison des problèmes de confidentialité des patients et des cliniciens. Les approches, par conséquent, qui peuvent adapter un modèle au domaine cible invisible et non étiqueté tel que la OR tout en préservant la confidentialité sont des poursuites prometteuses. Les approches d'adaptation de domaine non supervisé (UDA) visent à généraliser un modèle dans différents domaines avec accès aux données étiquetées du domaine source et aux données non étiquetées du domaine cible, voir l'exemple dans la Figure B.1. Cette thése explore plusieurs approches d'adaptation de domaine non supervisée (UDA) pour l'estimation HPE et la segmentation d'instances de personnes, comme indiqué ci-dessous.

B.3 Contributions

B.3.1 Publication de l'ensemble de données *MVOR* et comparaison des méthodes de pointe

Comme première contribution, nous introduisons un nouvel ensemble de données de salle d'opération à vues multiples, appelé *MVOR*, en tant que premier ensemble de données

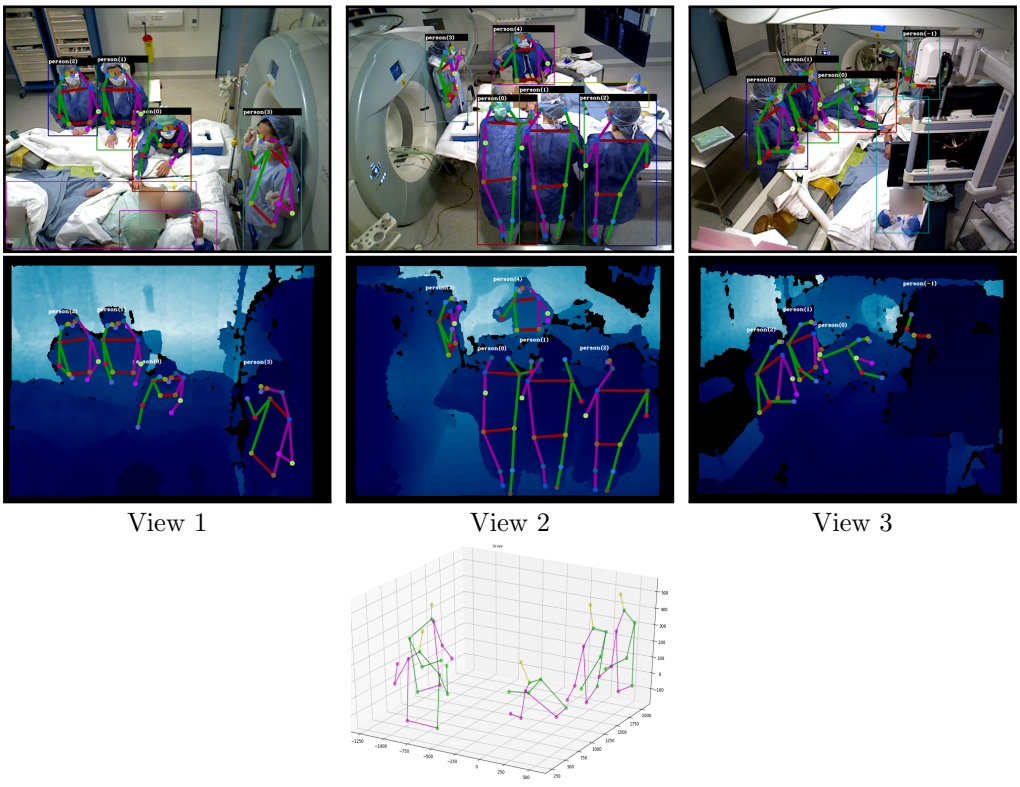


Figure B.3: exemples d'images et d'annotations de l'ensemble de données *MVOR*. L'ensemble de données et le code sont disponibles ici : https://github.com/CAMMA-public/MVOR

public enregistré lors d'interventions cliniques réelles pour la détection multi-personnes et 2D/3D, voir Figure B.2. *MVOR* montre les défis visuels inhérents à l'environnement réel de la salle d'opération, illustrant des variations significatives dans la distribution des couleurs par rapport aux images naturelles et aux cliniciens près les uns des autres portant des vêtements amples et des masques. Nous publions l'ensemble de données *MVOR* dans le but de faire progresser l'état de l'art pour la localisation précise des personnes dans la salle d'opération en le proposant comme un banc de test comparatif, voir la figure B.3. Nous évaluons des approches de pointe pour la détection de personnes et l'estimation de la pose humaine à l'échelle originale (1x) et à l'échelle sous-échantillonnée (12x). Nous observons une dégradation significative des performances, en particulier sur les images basse résolution, ce qui nous aide à établir des planchers initiaux *source-only* pour nos approches UDA.

B.3.2 Adaptation de domaine non supervisée à travers les modalités visuelles pour des images de profondeur basse résolution

Comme deuxième contribution, nous concevons des approches d'adaptation de domaine non supervisée pour l'estimation de pose humaine sur des images de profondeur basse résolution. Comme les images de profondeur sont sans texture et n'encodent que la

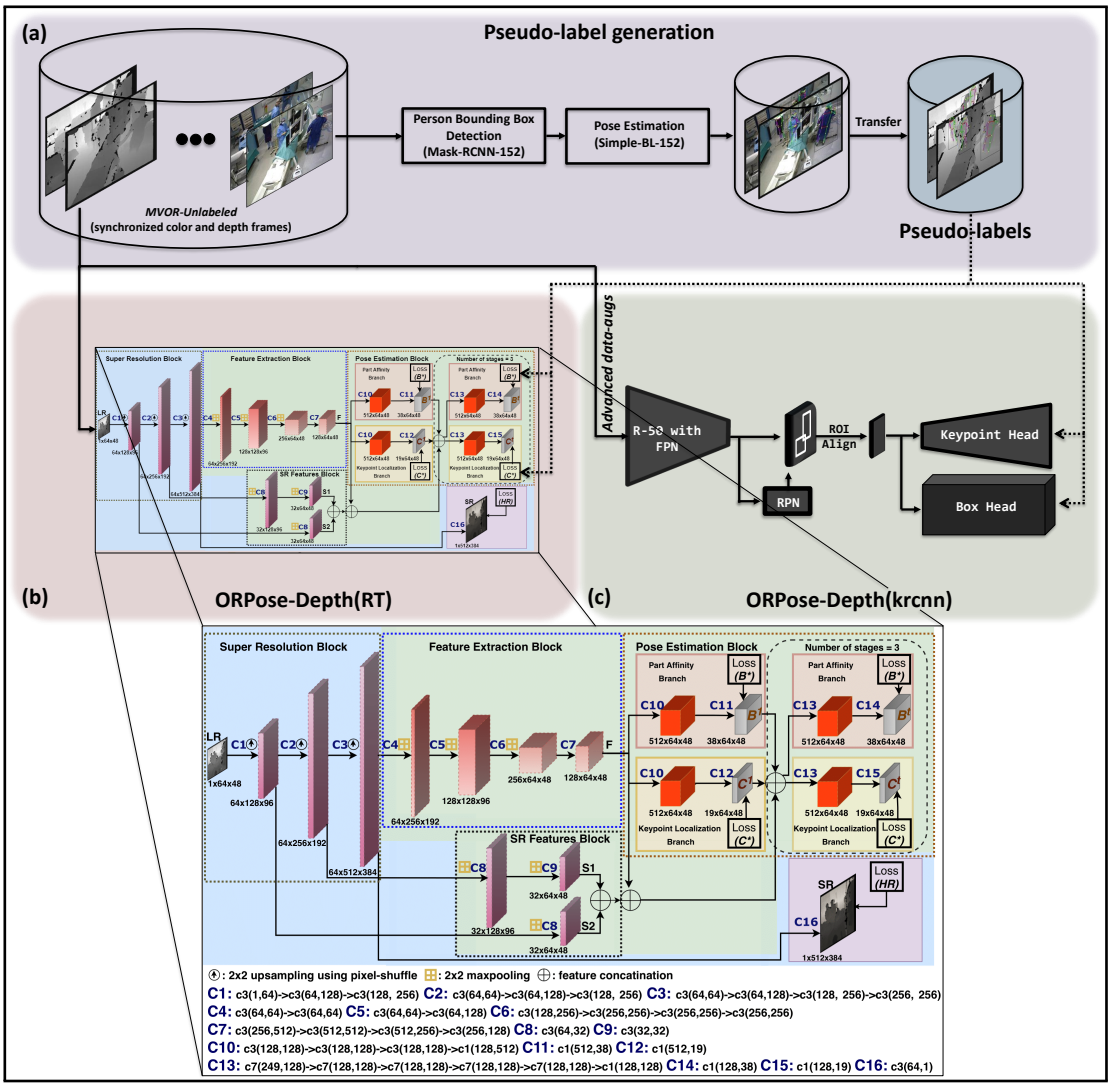


Figure B.4: Approche proposée. (a) Génération de pseudo-étiquettes : nous utilisons un détecteur de personnes à cadre (Mask-RCNN avec ResNet-152) suivi d'un estimateur de pose d'une seule personne (Simple-BL avec ResNet-152) pour générer les pseudo-étiquettes sur les images couleur de *MVOR-unlabeled*. Ces étiquettes sont ensuite transférées sur les images de profondeur correspondantes. (b) ORPose-Depth(RT) : nous proposons de modifier l'architecture ascendante RTPose pour entraîner le modèle sur des images de profondeur à faible résolution. Le bloc de super-résolution augmente la résolution spatiale d'un facteur 8 et génère des cartes de caractéristiques super-résolues intermédiaires (S1, S2) utilisées par le bloc d'estimation de pose pour apprendre des caractéristiques à haute fréquence. Toutes les pertes sont des pertes d'erreur quadratiques moyennes. C1 à C16 sont des couches de convolution regroupées pour une meilleure visualisation et décrites sous la figure, où c1(n1,n2), c3(n1,n2), c7(n1,n2) représentent chacune une couche de convolution avec une taille de noyau 1x1, 3x3, 7x7 et remplissage 0, 1, 3, respectivement. Les paramètres n1 et n2 sont les nombres de canaux d'entrée et de sortie, et toutes les couches de convolution sont suivies par la fonction non-linéaire RELU. (c) ORPose-Depth(krcnn) : nous proposons d'utiliser des augmentations de données avancées pour mieux apprendre les fonctionnalités à basse résolution dans le modèle Keypoint-RCNN descendant.

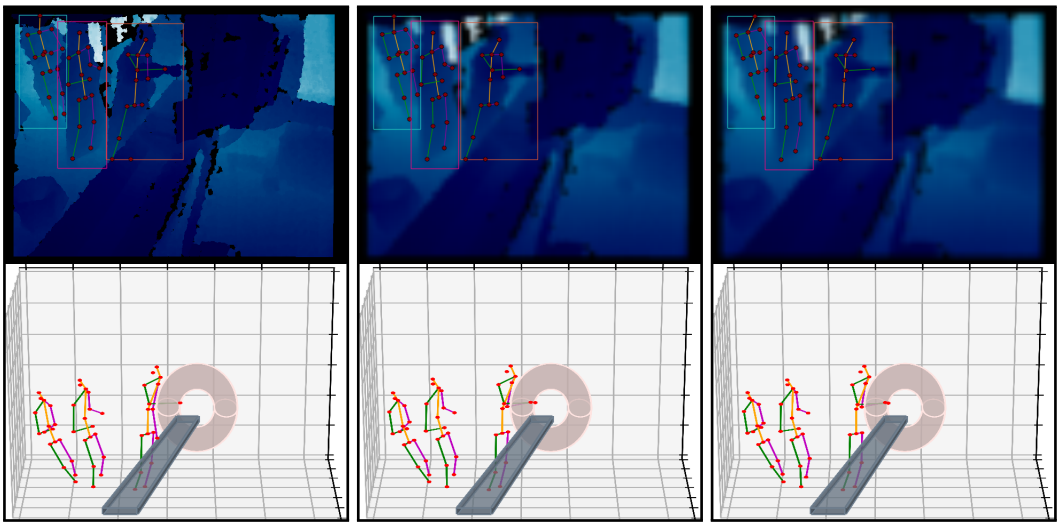


Figure B.5: Un exemple de résultat qualitatif de notre approche d'adaptation de domaine non supervisée sur différentes images de profondeur à basse résolution pour l'estimation de la pose humaine 2D/3D. La vidéo de démonstration est disponible ici : https://cutt.ly/depthpose. Page du projet : https://github.com/CAMMA-public/ORPose-depth

distance entre un objet et le capteur, elles offrent une option viable pour préserver la confidentialité des patients et des cliniciens. Comme souligné dans 1.2.1, nous n'utilisons en outre que les images de profondeur basse résolution au moment du test pour appliquer des contraintes plus importantes pour la salle d'opération, sensible à la confidentialité.

Pour entraîner un modèle sur les images de profondeur basse résolution sans annotations manuelles, nous avons avancé l'idée que deux modalités visuelles différentes, telles que les images couleur et profondeur, peuvent servir en tant que deux domaines distincts. Si ces deux domaines sont synchronisés, aussi simplement que possible grâce aux caméras RGBD, alors un modèle fonctionnant raisonnablement bien sur un domaine peut effectivement être adapté à l'autre. Pour ce faire, nous proposons d'effectuer une inférence sur les images couleur à l'aide de modèles d'estimation de pose humaine de pointe, d'affiner les résultats d'inférence pour générer des pseudo-étiquettes et de transférer les pseudo-étiquettes sur l'image de profondeur correspondante pour l'entraînement.

Nous proposons en outre deux stratégies d'entraînement utilisant les pseudo-étiquettes générées pour adapter un modèle aux images de profondeur à basse résolution. Comme première stratégie, nous proposons d'intégrer des cartes de caractéristiques super-résolues dans la méthode ascendante RTPose [Cao 2017] qui utilise des cartes de caractéristiques intermédiaires super-résolues pour un apprentissage efficace des caractéristiques à haute fréquence. Comme deuxième stratégie, nous exploitons des augmentations de données avancées telles que le sous-échantillonnage et l'échantillonnage à basse résolution, l'augmentation aléatoire [Cubuk 2020] et le recadrage aléatoire [DeVries 2017] dans le Keypoint-RCNN descendant [He 2017]. La figure B.4

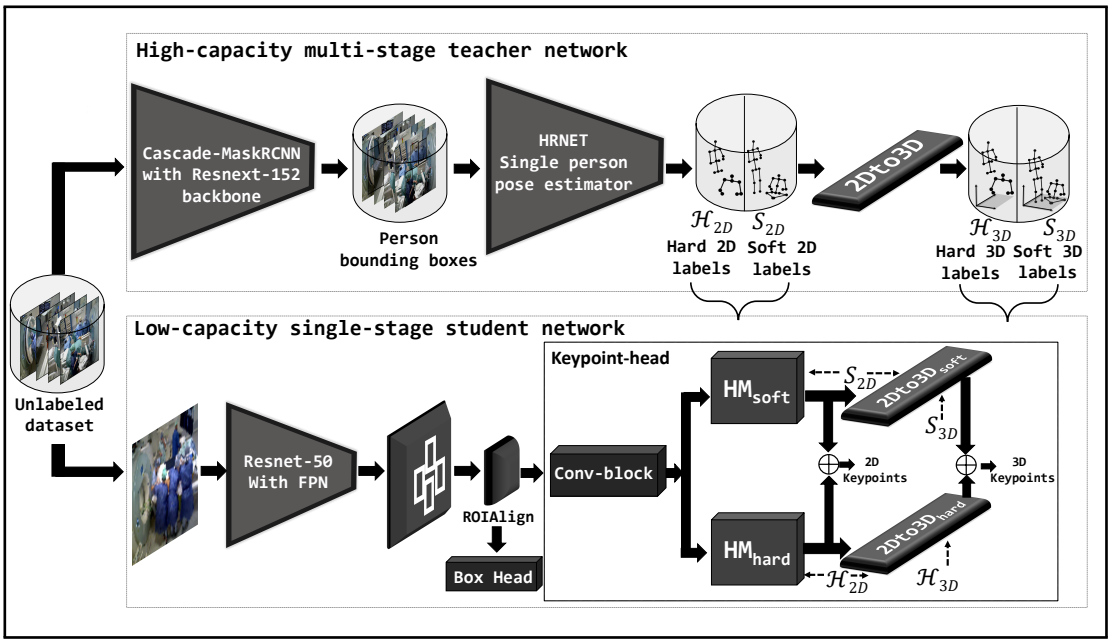


Figure B.6: Méthodologie auto-supervisée proposée pour l'estimation conjointe de points clés 2D/3D à l'aide du paradigme maître/élève. Le réseau maître est un réseau à trois étages qui utilise l'ensemble de données non étiqueté pour extraire les cadres de délimitation des personnes, estimer les points clés 2D et régresser les points clés 2D en 3D. Il génère des pseudo-étiquettes logicielles et matérielles à utiliser par le réseau élève. Le réseau élève est un réseau à un seul étage et utilise efficacement les pseudo-étiquettes, définies de façon souple ou dure, pour estimer conjointement les points clés 2D et 3D.

montre l'architecture de notre approche proposée. Nous montrons des résultats nettement meilleurs pour nos deux stratégies sur *MVOR*, ensemble de données difficile; en particulier sur les images de profondeur à basse résolution préservant la confidentialité avec un facteur de sous-échantillonnage allant jusqu'à 12x. La figure B.5 montre quelques résultats qualitatifs de notre approche.

B.3.3 Auto-supervision sur des images couleur OU non étiquetées pour l'estimation conjointe de la pose humaine 2D/3D

Pour notre troisième contribution, nous proposons une approche d'adaptation de domaine non supervisée fondée sur le paradigme d'apprentissage *enseignant-élève* pour développer un modèle facilement déployable pour l'estimation conjointe de la pose humaine 2D/3D sur les images couleur de la salle d'opération. Le modèle maître exploite la distillation des connaissances [Hinton 2015, Zhang 2019a] - en utilisant des modèles complexes à trois étages - ainsi que la distillation des données [Radosavovic 2018] pour générer des pseudo-étiquettes précises. Nous proposons d'utiliser deux ensembles d'étiquettes du modèle maître: un ensemble "dur" en supprimant les détections à faible confiance et un ensemble "souple" en conservant toutes les détections avec leur valeur de confiance.

Nous proposons en outre un modèle élève en une seule étape de bout en bout fondé sur

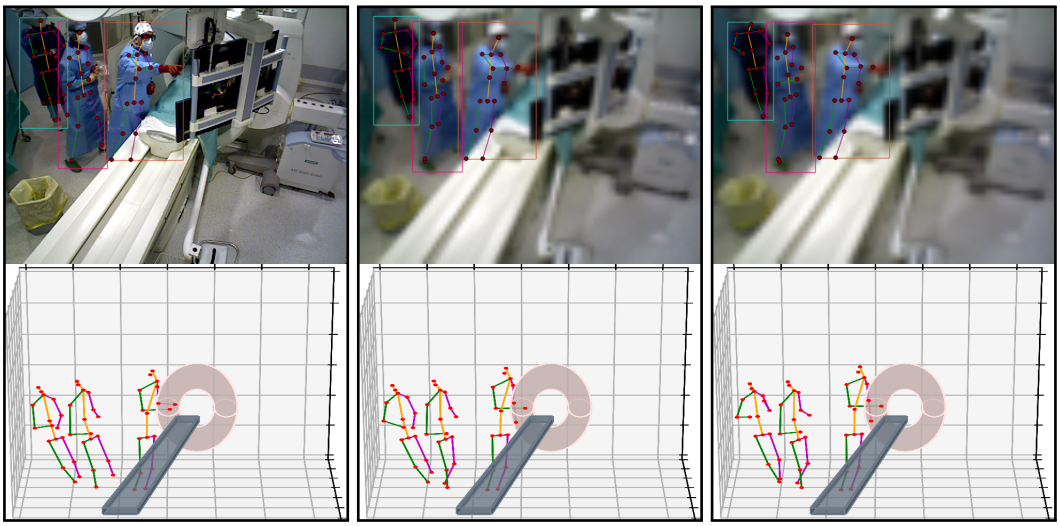


Figure B.7: Un exemple de résultat qualitatif de notre approche d'adaptation de domaine non supervisée sur différentes images couleur basse résolution pour l'estimation de pose humaine 2D/3D multi-personnes. La vidéo de démonstration est disponible ici : https://cutt.ly/orpose3d. Page du projet : https://github.com/CAMMA-public/ORPose-Color

Mask-RCNN [He 2017] où nous remplaçons la tête à masque par une tête à point clé pour l'estimation conjointe de pose 2D et 3D. Le modèle élève exploite à la fois les étiquettes "dures" et "souples" pour un entraînement efficace. De plus, pour adapter le modèle aux images basse résolution préservant la confidentialité, nous étendons le processus d'augmentation de données pour générer des images basse résolution en sous-échantillonnant et en sur-échantillonnant l'image d'entrée venant de la salle d'opération avec un facteur d'échelle aléatoire compris entre 1 et 12. Le modèle entraîné sur ces images de salle d'opération à très basse résolution apprend à donner des résultats précis au fur et à mesure de la progression de la formation. La figure B.6 montre l'architecture de notre approche proposée. Les résultats sur *MVOR* montrent que le modèle élève égale le réseau maître, bien qu'il soit léger et à un seul étage. De plus, il peut également donner des résultats précis sur des images à basse résolution, comme nécessaire pour garantir la confidentialité, même en utilisant un taux de sous-échantillonnage de 12x. La figure B.7 montre quelques résultats qualitatifs de notre approche.

B.3.4 Adaptation de domaine non supervisée pour l'estimation de la pose du clinicien et la segmentation des instances dans la salle d'opération

Dans nos deux premières contributions, nous proposons d'utiliser un modèle maître multi-étapes robuste pour générer de façon précise des pseudo-étiquettes. Cependant, un modèle maître peut ne pas toujours être disponible pour entraîner un modèle élève. Dans ce travail, nous posons la question suivante : au lieu de s'appuyer sur un modèle

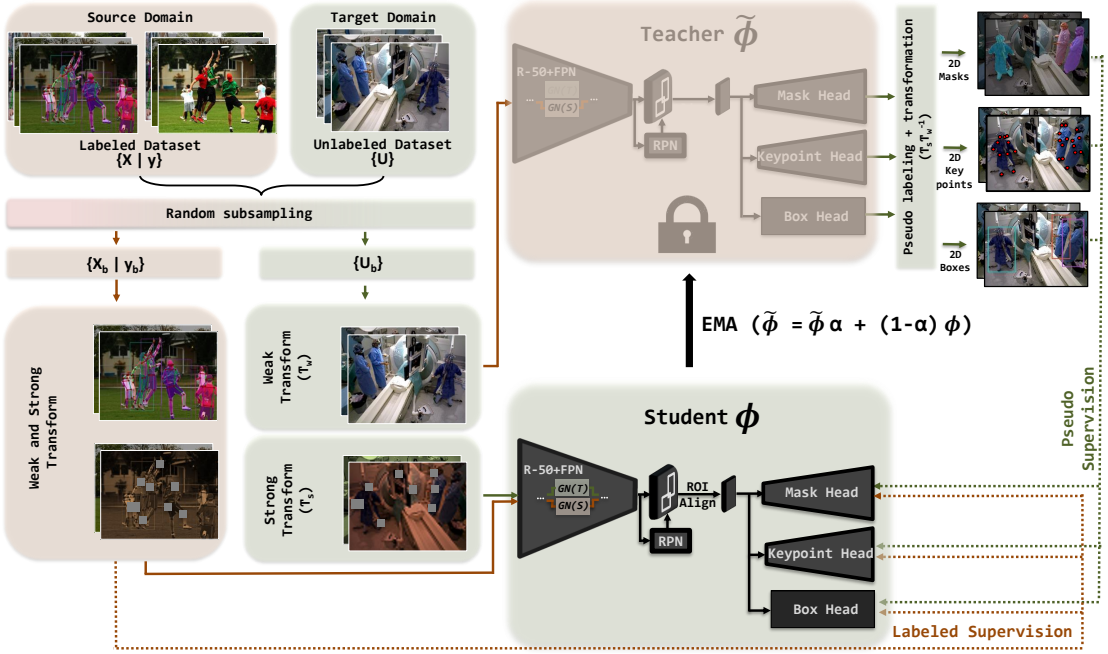
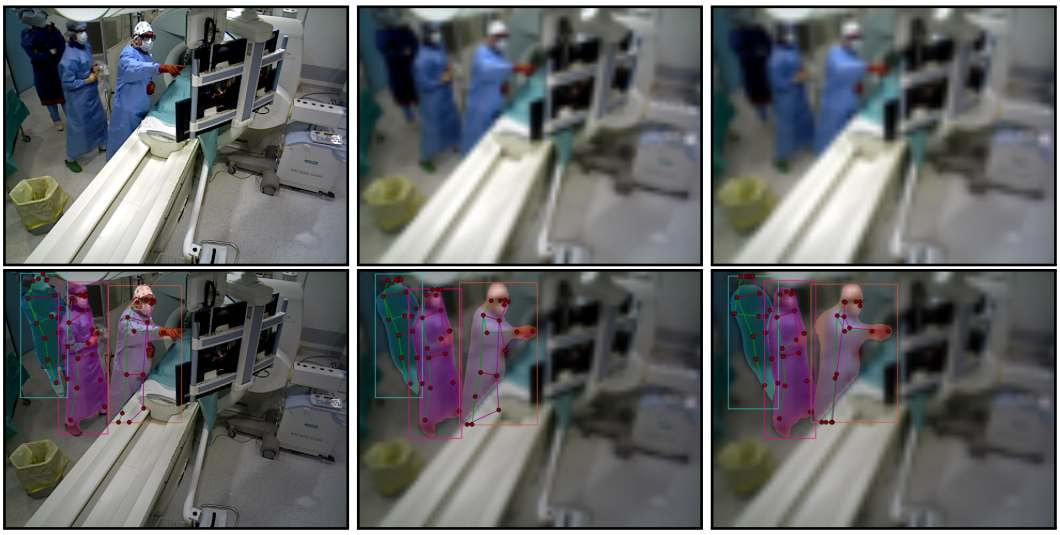


Figure B.8: Présentation de notre approche pour l'adaptation de domaine non supervisée. Nous générons deux types d'augmentations sur les images de domaine cible non étiquetées : faible et forte. Les images faiblement augmentées passent par un modèle maître gelé et une fonction de seuillage pour générer les pseudo-étiquettes. Ces pseudo-étiquettes sont ensuite transformées géométriquement dans l'espace image fortement augmenté correspondant. Un modèle élève utilise ces pseudo-étiquettes transformées pour s'entraîner sur les images non étiquetées fortement augmentées conjointement avec les images du domaine source étiquetées. Les poids du modèle d'enseignant gelé sont misà jour à l'aide de la moyenne mobile exponentielle (EMA) des poids du modèle d'élève. Nous remplaçons également chaque couche de normalisation de groupe (GN) dans l'extracteur de caractéristiques par deux couches GN (*GN(S)* et *GN(T)*) pour normaliser les caractéristiques des deux domaines séparément, selon les besoins pour gérer les domaines source et cible, statistiquement différents.

maître fort pour donner des pseudo-étiquettes, un modèle peut-il devenir son propre maître pour l'entraînement?

Comme dernière contribution, nous proposons une nouvelle approche d'adaptation de domaine non supervisée, appelée *AdaptOR*, pour l'estimation de la pose d'une personne conjointement à la segmentation des instances. Nous proposons d'abord d'étendre le Mask R-CNN [He 2017] en utilisant DFN pour s'entraîner sur deux domaines statistiquement différents : les images naturelles de *COCO*, et les images OR de MVOR ou TUM-OR [Belagiannis 2016]. DFN remplace chaque couche de normalisation de caractéristiques dans l'armature extractrice de caractéristiques par deux couches de normalisation: une pour le domaine source et une autre pour le domaine cible. Nous proposons de modifier la fonction de perte pour améliorer l'armature. Nous transmettons les caractéristiques des domaines source et cible séparément aux têtes en aval, nécessaires pour pondérer séparément les pertes pour les deux domaines.

Étant donné une armature avec la capacité de s'entraîner sur deux domaines



1x 8x 10x

Figure B.9: Un exemple de résultat qualitatif de notre approche d'adaptation de domaine non supervisée sur différentes images couleur basse résolution pour l'estimation de la pose d'une personne conjointement à la segmentation des instances. Vidéo de démonstration : https://youtu.be/gqwPu9-nfGs, page du projet : https://github.com/CAMMA-public/HPE-AdaptOR

statistiquement distincts, nous proposons d'exploiter des contraintes géométriques explicites sur les différentes augmentations des images du domaine cible non étiquetées pour générer de manière précise des pseudo-étiquettes et d'utiliser ces pseudo-étiquettes pour entraîner le modèle sur et des images OR à basse résolution. Les contraintes géométriques doivent satisfaire des *contraintes équivariantes de transformation* en transformant les prédictions du modèle pour observer les mêmes augmentations géométriques que sur l'image d'entrée. Ces contraintes géométriques explicites aident le modèle à s'adapter efficacement au domaine cible. Nous étendons encore notre approche avec un "maître moyen" pour un entraînement stable [Tarvainen 2017], où au lieu d'utiliser un modèle unique pour générer et consommer les pseudo-étiquettes, nous créons deux copies d'un modèle entraîné sur le domaine source: un *maître* et un modèle *élève*. Le modèle *maître* génère les pseudo-étiquettes sur l'image *faiblement* augmentée qui est utilisée par le modèle *élève* pour s'entraîner sur l'image *fortement* augmentée correspondante. Les poids du modèle *maître* sont mis à jour à l'aide d'un assemblage temporel des poids du modèle *élève*, l'aidant ainsi à améliorer ses prédictions grâce à l'assemblage tout en générant simultanément de meilleures pseudo-étiquettes pour améliorer le modèle *élève*. La figure B.8 montre l'architecture de notre approche.

L'évaluation de la méthode sur les deux ensembles de données du domaine cible, *MVOR* et *TUM-OR*, avec des études d'ablation approfondies, montre l'efficacité de notre approche. Les résultats nettement meilleurs sur les images à basse résolution encouragent l'utilisation de notre méthode pour la salle d'opération, sensible à la vie privée. La figure B.9 montre quelques résultats qualitatifs de notre approche sur un exemple d'image *MVOR*.

B.4 Conclusion

La localisation précise des cliniciens en salle d'opération est un élément clé dans la conception de la nouvelle génération de systèmes d'assistance au bloc opératoire. Le développement de modèles de vision par ordinateur pour la segmentation d'images basée sur les pixels ainsi que la détection des points saillants du corps est nécessaire pour mieux comprendre les activités cliniques et la disposition spatiale de la salle d'opération. Cette tâche, cependant, pose des défis dûs aux deux limitations suivantes : 1) différence significative de domaine visuel entre les images du bloc opératoire et celle de l'ensemble des données de vision traditionnelle, et 2) obstacles à la confidentialité lors de la collecte et de la génération des données et des annotations.

L'objectif principal de cette thèse est de développer des approches exploitant des données de salle d'opération non annotées à grande échelle pour une localisation précise des personnes tout en préservant la confidentialité des cliniciens. Cette thèse explore les méthodes d'adaptation de domaine non supervisée (UDA pour Unsupervised Domain Adaptation) afin d'améliorer l'apprentissage visuel pour le domaine cible, le bloc opératoire, et ce en travaillant dans deux directions complémentaires. Tout d'abord, nous étudions comment des images basse résolution avec un facteur de sous-échantillonnage aussi faible que 12x peuvent être utilisées pour une localisation précise des cliniciens, ce sous-échantillonage permettant de résoudre les problèmes de confidentialité. Deuxièmement, nous proposons plusieurs méthodes auto-supervisées pour transférer les informations apprises d'un domaine source annoté vers un domaine cible non annoté pour faire face au changement de domaine visuel et au manque d'annotations. Ces méthodes utilisent des prédictions auto-supervisées pour permettre au modèle d'apprendre et de s'adapter au domaine cible non annoté. Nous obtenons des résultats qui font état de l'art sur l'ensemble des images du jeu de données MVOR recueilli à partir d'interventions cliniques réelles. Nous espérons que cette thèse pourra contribuer à la mise à l'échelle et au déploiement de nouvelles applications d'assistance utilisant l'IA pour la salle d'opération.

B.5 Perspectives

Nous pensons que plusieurs axes de recherche pourraient naître des travaux présentés dans cette thèse exploitant de manière novatrice des données non étiquetées à grande échelle. Ceux-ci pourraient ouvrir des possibilités pour explorer des problèmes de recherche novateurs, permettant ainsi au bloc opératoire du futur de devenir le bloc opératoire du présent. Différentes perspectives de notre travail sont présentées ci-dessous.

B.5.1 Estimation de pose 3D absolue en multi-vues et multi-personnes

Dans cette thèse, nous avons proposé des approches de localisation de personne en vue unique utilisant uniquement les données non annotées provenant de la salle d'opération. Cependant, avec une seule vue, il est possible que les cliniciens ne soient pas visibles en

raison de fortes occlusions. Cela pourrait éventuellement entraîner de mauvaises performances, en particulier dans l'estimation de pose 3D, qui repose sur une donnée d'entrée précise de l'estimation de pose 2D, voir la figure B.10. Les images multi-vues peuvent non seulement aider à résoudre les cas d'occlusion, mais également aider à fournir une estimation de pose 3D absolue dans le repère de coordonnées de la salle d'opération, par opposition aux coordonnées relatives à la racine principalement utilisées dans l'estimation de pose 3D pour une seule personne. La recherche actuelle sur l'estimation de pose 3D absolue en multi-vues et multi-personnes est cependant limitée à des approches entièrement supervisées [Tu 2020, Reddy 2021] sur des ensembles de données acquis lors d'activités sociales simulées [Joo 2015]. L'utilisation de l'adaptation de domaine non supervisé gls uda permettrait d'étendre ces approches entièrement supervisées à un environnement de salle d'opération réel non annoté, avec un nombre de perspectives limité. Plus intéressant encore, une autre direction prometteuse serait d'apprendre l'estimation de pose 3D absolue de manière purement auto-supervisée. Lorsque des images multi-vues sont disponibles, la pose 3D en coordonnées absolues peut être déterminée en utilisant des approches de géométrie multi-vues et de

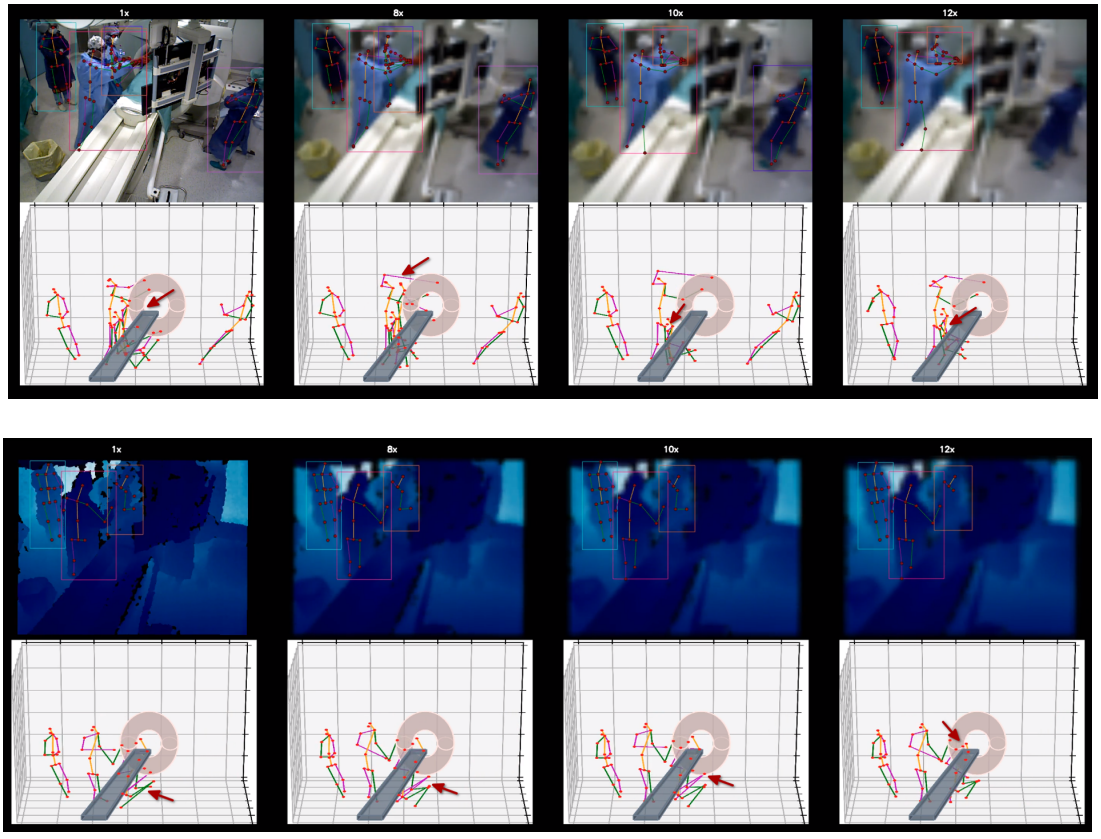


Figure B.10: Certains des cas d'échec de nos approches d'estimation de pose 3D sur les images couleur et les images profondeur sont principalement dus à la forte occlusion des autres cliniciens et à l'encombrement des instruments. Les images multi-vues peuvent aider à résoudre ces situations d'échec en prenant en compte des informations complémentaires à partir d'autres vues.

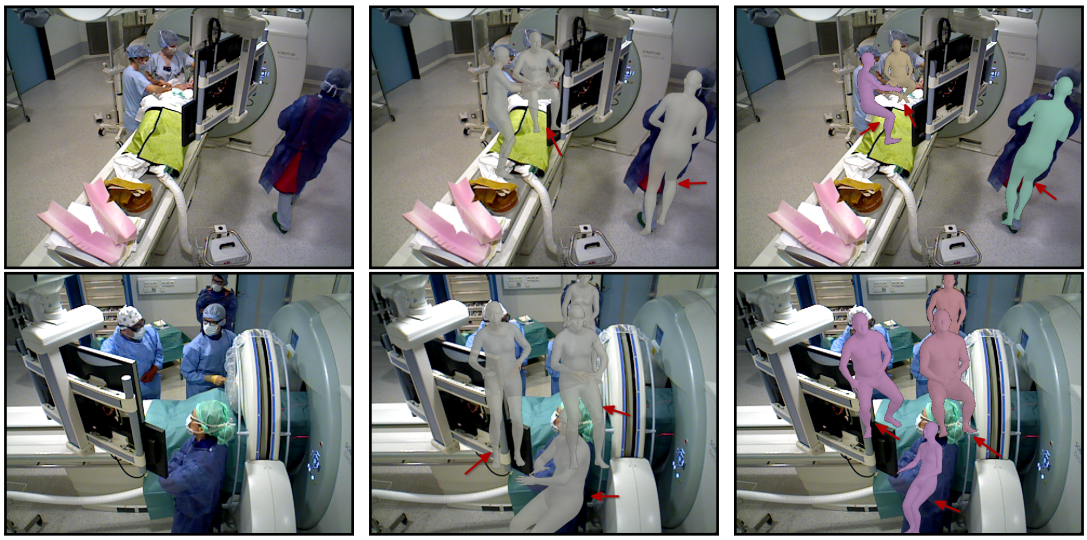
triangulation [Andrew 2001]. Cependant, un obstacle important réside dans l'imprécision des estimations de pose 2D en raison des potentielles occlusions et des vues de caméra limitées. Les contraintes spatiales et géométriques décrites au chapitre 6 peuvent être étendues dans l'espace volumétrique multi-vues pour affiner de manière itérative les poses 3D en exploitant ces contraintes, la triangulation ainsi qu'une estimation plus robuste de l'estimation de pose humaine en 2D.

B.5.2 Localisation de personnes multi-modalités à l'aide d'images RGBD

Cette thèse s'est principalement concentrée sur l'utilisation de données non annotées pour développer diverses approches fines de localisation de personnes. Cependant, les approches développées prennent des images d'entrée à modalité unique (couleur ou profondeur) au moment de l'inférence. Avec l'avènement des caméras RGBD peu coûteuses, les images de profondeur et de couleur sont facilement disponibles, fournissant des informations complémentaires sur la scène - des informations sur la texture 2D à partir de la couleur et des informations sur l'environnement spatial 3D à partir des images de profondeur. Les modèles basés sur des réseaux convolutionnels CNN (Convolutional Neural Networks) ont montré des performances nettement améliorées pour les images RGB. Cependant, l'application directe de ces modèles CNN sur des images de profondeur peut être sous-optimale. Une direction de recherche intéressante serait d'explorer le développement d'une architecture unifiée pour les images RGBD exploitant les informations complémentaires des deux modalités au moment de l'inférence. Une telle architecture peut aider non seulement à améliorer la précision au moment de l'inférence, mais également à étendre l'approche UDA, comme discuté au chapitre 6, pour une adaptation de domaine plus robuste.

B.5.3 Exploitation de la temporalité pour une estimation précise de la pose et un suivi cohérent

La temporalité est une autre dimension qui peut être exploitée pour une stabilité et une précision accrue de la localisation. Par exemple, dans le cadre des approches d'adaptation de domaine non-supervisé UDA basées sur les modèles *enseignant-élève* décrit dans le chapitre 5, on peut exploiter la temporalité dans le modèle de l'enseignant pour obtenir des pseudo-étiquettes plus précises. Le modèle étudiant peut lui aussi être étendu avec la composante temporelle pour exploiter ces pseudo-étiquettes précises afin d'obtenir une donnée de sortie plus précise et un suivi plus cohérent des cliniciens. Par ailleurs, les approches UDA décrites dans le chapitre 6 peuvent également exploiter la temporalité pour générer et utiliser plus précisément les pseudo étiquettes dans leur temporalité.



Input CenterHMR [Sun 2020] VIBE [Kocabas 2020]

Figure B.11: Echecs de certaines approches de pointe pour l'estimation de pose et de forme du corps humain pour des images provenant de l'environnement exigeant qu'est le bloc opératoire

B.5.4 Estimation de la pose et de la forme du corps humain

Cette thèse s'est principalement concentré sur la localisation fine de la personne allant de la segmentation d'instances à l'estimation de pose humaine en 2D et 3D. Bien qu'extrêmement utile, la segmentation d'instance est limitée à l'image 2D; l'estimation de pose, elle, localise les points saillants dispersés en 2D ou en 3D. Les directions de recherche actuelles ont évolué pour estimer le modèle de maillage humain 3D paramétrique à partir d'une image donnée, combinant ainsi la segmentation d'instance basée sur les pixels et l'estimation de pose 3D dans un champ unifié [Loper 2015, Kanazawa 2018]. Cependant, la recherche s'est principalement concentrée sur l'apprentissage du modèle de forme corporelle à partir d'ensembles de données synthétiques tels que Human3.6 [Ionescu 2013]. Lorsqu'elles sont appliquées à des environnements réels difficiles tels que le bloc opératoire OR, ces approches échouent remarquablement, comme illustré dans la figure B.11. Les approches d'adaptation de domaine non supervisées UDA discutées dans cette thèse pourraient aider à intégrer l'estimation de forme du corps humain au bloc opératoire OR.

References

[5Gw] *5G: what are the benefits that come from the hand of this new standard that allows speeds up to 100 times higher than 4G - QoriLab*. https://qorilab.com/5g-cuales-son-los-beneficios-que-llegan-de-la-mano-de-este-nuevo-estandar-que-permite-velocidades-l (Accessed on 10/01/2021). (Cited on pages xiii and 7)

[Agarwal 2007] S. Agarwal, A. Joshi, T. Finin, Y. Yesha and T. Ganous. *A pervasive computing system for the operating room of the future*. Mobile Networks and Applications, vol. 12, no. 2, pages 215–228, 2007. (Cited on page 9)

[Andrew 2001] A. M. Andrew. *Multiple view geometry in computer vision*. Kybernetes, 2001. (Cited on pages 107 and 141)

[Andriluka 2014] M. Andriluka, L. Pishchulin, P. Gehler and B. Schiele. *2d human pose estimation: New benchmark and state of the art analysis*. In Proceedings of the IEEE Conference on computer Vision and Pattern Recognition, pages 3686–3693, 2014. (Cited on pages xiv, 32, 37, 48, and 115)

[Arnab 2017] A. Arnab and P. H. Torr. *Pixelwise instance segmentation with a dynamically instantiated network*. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 441–450, 2017. (Cited on page 26)

[Bai 2017] M. Bai and R. Urtasun. *Deep watershed transform for instance segmentation*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 5221–5229, 2017. (Cited on page 26)

[Bai 2018] Y. Bai, Y. Zhang, M. Ding and B. Ghanem. *Finding tiny faces in the wild with generative adversarial network*. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 21–30, 2018. (Cited on page 22)

[Bardram 2011] J. E. Bardram, A. Doryab, R. M. Jensen, P. M. Lange, K. L. Nielsen and S. T. Petersen. *Phase recognition during surgical procedures using embedded and body-worn sensors*. In 2011 IEEE international conference on pervasive computing and communications (PerCom), pages 45–53. IEEE, 2011. (Cited on page 9)

[Bekhtaoui 2020] W. Bekhtaoui, R. Sa, B. Teixeira, V. Singh, K. Kirchberg, Y.-j. Chang and A. Kapoor. *View Invariant Human Body Detection and Pose Estimation from Multiple Depth Sensors.* arXiv preprint arXiv:2005.04258, 2020. (Cited on page 26)

[Belagiannis 2016] V. Belagiannis, X. Wang, H. B. B. Shitrit, K. Hashimoto, R. Stauder, Y. Aoki, M. Kranzfelder, A. Schneider, P. Fua, S. Ilic*et al.* *Parsing human skeletons in an operating room.* Machine Vision and Applications, vol. 27, no. 7, pages 1035–1046, 2016. (Cited on pages 14, 25, 32, 72, 80, 115, 130, and 137)

[Ben-David 2010] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira and J. W. Vaughan. *A theory of learning from different domains.* Machine learning, vol. 79, no. 1, pages 151–175, 2010. (Cited on page 20)

[Berthelot 2019a] D. Berthelot, N. Carlini, E. D. Cubuk, A. Kurakin, K. Sohn, H. Zhang and C. Raffel. *Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring.* arXiv preprint arXiv:1911.09785, 2019. (Cited on page 71)

[Berthelot 2019b] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver and C. Raffel. *Mixmatch: A holistic approach to semi-supervised learning.* arXiv preprint arXiv:1905.02249, 2019. (Cited on page 21)

[Bouget 2015] D. Bouget, R. Benenson, M. Omran, L. Riffaud, B. Schiele and P. Jannin. *Detecting surgical tools by modelling local appearance and global shape.* IEEE transactions on medical imaging, vol. 34, no. 12, pages 2603–2617, 2015. (Cited on page 9)

[Bridges 1999] M. Bridges and D. L. Diamond. *The financial impact of teaching surgical residents in the operating room.* The American Journal of Surgery, vol. 177, no. 1, pages 28–32, 1999. (Cited on page 10)

[Cai 2019a] Q. Cai, Y. Pan, C.-W. Ngo, X. Tian, L. Duan and T. Yao. *Exploring object relation in mean teacher for cross-domain detection.* In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 11457–11466, 2019. (Cited on page 21)

[Cai 2019b] Z. Cai and N. Vasconcelos. *Cascade R-CNN: high quality object detection and instance segmentation.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019. (Cited on pages 27, 39, 62, and 80)

[Cao 2017] Z. Cao, T. Simon, S.-E. Wei and Y. Sheikh. *Realtime multi-person 2d pose estimation using part affinity fields.* In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 7291–7299, 2017. (Cited on pages xiii, 5, 6, 13, 24, 25, 27, 33, 39, 40, 49, 50, 51, 52, 60, 115, 117, 118, 122, and 134)

[Carinou 2011] E. Carinou, M. Brodecki, J. Domienik, L. Donadille, C. Koukorava, S. Krim, D. Nikodemova, N. Ruiz-Lopez, M. Sans-Merce, L. Struelens*et al. Recommendations to reduce extremity and eye lens doses in interventional radiology and cardiology.* Radiation measurements, vol. 46, no. 11, pages 1324–1329, 2011. (Cited on page 10)

[Catchpole 2007] K. R. Catchpole, M. R. De Leval, A. McEwan, N. Pigott, M. J. Elliott, A. McQuillan, C. Macdonald and A. J. Goldman. *Patient handover from surgery to intensive care: using Formula 1 pit-stop and aviation models to improve safety and quality.* Pediatric anesthesia, vol. 17, no. 5, pages 470–478, 2007. (Cited on page 9)

[Chang 2019] W.-G. Chang, T. You, S. Seo, S. Kwak and B. Han. *Domain-specific batch normalization for unsupervised domain adaptation.* In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 7354–7362, 2019. (Cited on pages 21, 71, and 75)

[Chen 2017] C.-H. Chen and D. Ramanan. *3d human pose estimation= 2d pose estimation+ matching.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 7035–7043, 2017. (Cited on page 25)

[Chen 2018a] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu and J. Sun. *Cascaded pyramid network for multi-person pose estimation.* pages 7103–7112, 2018. (Cited on pages xiii, 6, 24, and 115)

[Chen 2018b] Y. Chen, W. Li, C. Sakaridis, D. Dai and L. Van Gool. *Domain adaptive faster r-cnn for object detection in the wild.* In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3339–3348, 2018. (Cited on page 20)

[Chen 2019a] C. Chen, Q. Dou, H. Chen, J. Qin and P.-A. Heng. *Synergistic image and feature adaptation: Towards cross-modality domain adaptation for medical image segmentation.* In Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, pages 865–872, 2019. (Cited on pages 20 and 22)

[Chen 2019b] X. Chen, R. Girshick, K. He and P. Dollár. *Tensormask: A Foundation for Dense Object Segmentation.* 2019. (Cited on page 26)

[Chen 2019c] Y. Chen, W. Li, X. Chen and L. V. Gool. *Learning semantic segmentation from synthetic data: A geometrically guided input-output adaptation approach.* In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 1841–1850, 2019. (Cited on page 20)

[Chen 2019d] Y.-C. Chen, Y.-Y. Lin, M.-H. Yang and J.-B. Huang. *Crdoco: Pixel-level domain transfer with cross-domain consistency.* In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 1791–1800, 2019. (Cited on page 20)

[Chen 2020] T. Chen, S. Kornblith, M. Norouzi and G. Hinton. *A simple framework for contrastive learning of visual representations.* In International conference on machine learning, pages 1597–1607. PMLR, 2020. (Cited on pages xxii, 12, 75, 81, and 83)

[Chen 2021] Z. Chen, X. Guo, P. Y. Woo and Y. Yuan. *Super-Resolution Enhanced Medical Image Diagnosis With Sample Affinity Interaction.* IEEE Transactions on Medical Imaging, vol. 40, no. 5, pages 1377–1389, 2021. (Cited on page 23)

[Cheng 2017] Z. Cheng, T. Shi, W. Cui, Y. Dong and X. Fang. *3D face recognition based on kinect depth data.* In 4th International Conference on Systems and Informatics (ICSAI), pages 555–559, 2017. (Cited on page 48)

[Cheng 2020] B. Cheng, B. Xiao, J. Wang, H. Shi, T. S. Huang and L. Zhang. *Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation.* In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5386–5395, 2020. (Cited on page 24)

[Childers 2018] C. P. Childers and M. Maggard-Gibbons. *Understanding costs of care in the operating room.* JAMA surgery, vol. 153, no. 4, pages e176233–e176233, 2018. (Cited on page 7)

[Choi 2019] J. Choi, T. Kim and C. Kim. *Self-ensembling with gan-based data augmentation for domain adaptation in semantic segmentation.* In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 6830–6840, 2019. (Cited on page 20)

[Chou 2018] E. Chou, M. Tan, C. Zou, M. Guo, A. Haque, A. Milstein and L. Fei-Fei. *Privacy-Preserving Action Recognition for Smart Hospitals using Low-Resolution Depth Images.* NeurIPS Workshop on Machine Learning for Health (ML4H), 2018. (Cited on pages 8, 22, 23, and 48)

[Cordts 2016] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth and B. Schiele. *The cityscapes dataset for semantic urban scene understanding.* In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3213–3223, 2016. (Cited on page 5)

[Csurka 2017] G. Csurka. *Domain adaptation for visual applications: A comprehensive survey.* arXiv preprint arXiv:1702.05374, 2017. (Cited on pages xiv and 19)

[Cubuk 2019] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan and Q. V. Le. *Autoaugment: Learning augmentation strategies from data.* In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 113–123, 2019. (Cited on page 78)

[Cubuk 2020] E. D. Cubuk, B. Zoph, J. Shlens and Q. V. Le. *Randaugment: Practical automated data augmentation with a reduced search space.* In Proceedings of the

IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pages 702–703, 2020. (Cited on pages 13, 49, 53, 78, 82, and 134)

[Dabral 2019] R. Dabral, N. B. Gundavarapu, R. Mitra, A. Sharma, G. Ramakrishnan and A. Jain. *Multi-Person 3D Human Pose Estimation from Monocular Images.* In 2019 International Conference on 3D Vision (3DV), pages 405–414. IEEE, 2019. (Cited on pages 25 and 61)

[Dai 2017] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu and Y. Wei. *Deformable convolutional networks.* In Proceedings of the IEEE international conference on computer vision, pages 764–773, 2017. (Cited on page 80)

[Daumé III 2009] H. Daumé III. *Frustratingly easy domain adaptation.* arXiv preprint arXiv:0907.1815, 2009. (Cited on page 20)

[Deng 2009] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei. *Imagenet: A large-scale hierarchical image database.* In 2009 IEEE conference on computer vision and pattern recognition, pages 248–255. Ieee, 2009. (Cited on page 75)

[Deng 2021] J. Deng, W. Li, Y. Chen and L. Duan. *Unbiased Mean Teacher for Cross-Domain Object Detection.* In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 4091–4101, 2021. (Cited on pages 21 and 71)

[DeVries 2017] T. DeVries and G. W. Taylor. *Improved regularization of convolutional neural networks with cutout.* arXiv preprint arXiv:1708.04552, 2017. (Cited on pages 13, 49, 53, 78, 82, and 134)

[Dias 2019] R. D. Dias, M. A. Zenati, R. Stevens, J. M. Gabany and S. J. Yule. *Physiological synchronization and entropy as measures of team cognitive load.* Journal of biomedical informatics, vol. 96, page 103250, 2019. (Cited on page 11)

[DiPietro 2019] R. DiPietro and G. D. Hager. *Automated surgical activity recognition with one labeled sequence.* In International conference on medical image computing and computer-assisted intervention, pages 458–466. Springer, 2019. (Cited on page 22)

[Donaldson 2000] M. S. Donaldson, J. M. Corrigan, L. T. Kohn*et al. To err is human: building a safer health system.* 2000. (Cited on page 10)

[Dong 2020] J. Dong, Y. Cong, G. Sun, B. Zhong and X. Xu. *What can be transferred: Unsupervised domain adaptation for endoscopic lesions segmentation.* In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 4023–4032, 2020. (Cited on page 22)

[Dou 2021] Q. Dou, T. Y. So, M. Jiang, Q. Liu, V. Vardhanabhuti, G. Kaissis, Z. Li, W. Si, H. H. Lee, K. Yu*et al. Federated deep learning for detecting COVID-19*

lung abnormalities in CT: a privacy-preserving multinational validation study. NPJ digital medicine, vol. 4, no. 1, pages 1–11, 2021. (Cited on page 22)

[Du 2019] L. Du, J. Tan, H. Yang, J. Feng, X. Xue, Q. Zheng, X. Ye and X. Zhang. *Ssf-dan: Separated semantic feature based domain adaptation network for semantic segmentation.* In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 982–991, 2019. (Cited on page 20)

[Duhaime] D. Duhaime, P. Leonard, T. Eskildsen, S. Choudhary, C. DeRose, W. Sanger, D. Reagan and o. Sorba. *YaleDHLab/pix-plot: A WebGL viewer for UMAP or TSNE-clustered images.* https://github.com/YaleDHLab/pix-plot. (Accessed on 07/30/2021). (Cited on pages xiv and 12)

[Efros 2019] A. Efros. *Alexei Efros: In the end, it's all about the Data.* https://youtu.be/M1VHu1d4sGQ?t=2339, 2019. The story appears at 39:00. (Cited on page 5)

[Fang 2017] H.-S. Fang, S. Xie, Y.-W. Tai and C. Lu. *Rmpe: Regional multi-person pose estimation.* In Proceedings of the IEEE International Conference on Computer Vision, pages 2334–2343, 2017. (Cited on pages xiii, 6, 24, 115, 118, and 122)

[Felzenszwalb 2005] P. F. Felzenszwalb and D. P. Huttenlocher. *Pictorial structures for object recognition.* International journal of computer vision, vol. 61, no. 1, pages 55–79, 2005. (Cited on page 25)

[Fischler 1973] M. A. Fischler and R. A. Elschlager. *The representation and matching of pictorial structures.* IEEE Transactions on computers, vol. 100, no. 1, pages 67–92, 1973. (Cited on pages 24 and 25)

[Flouty 2018] E. Flouty, O. Zisimopoulos and D. Stoyanov. *Faceoff: anonymizing videos in the operating rooms.* In OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis, pages 30–38. Springer, 2018. (Cited on page 115)

[Friedman 2000] J. Friedman, T. Hastie, R. Tibshirani*et al. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors).* The annals of statistics, vol. 28, no. 2, pages 337–407, 2000. (Cited on page 115)

[Gawande 2011] A. Gawande. *The checklist manifesto: How to get things right.* Journal of Nursing Regulation, vol. 1, no. 4, page 64, 2011. (Cited on page 9)

[Ge 2018] S. Ge, S. Zhao, C. Li and J. Li. *Low-resolution face recognition in the wild via selective knowledge distillation.* IEEE Transactions on Image Processing, vol. 28, no. 4, pages 2051–2062, 2018. (Cited on page 22)

[Ghani 2016] K. R. Ghani, D. C. Miller, S. Linsell, A. Brachulis, B. Lane, R. Sarle, D. Dalela, M. Menon, B. Comstock, T. S. Lendvay*et al. Measuring to improve: peer and crowd-sourced assessments of technical skill with robot-assisted radical*

prostatectomy. European urology, vol. 69, no. 4, pages 547–550, 2016. (Cited on page 10)

[Girshick 2014] R. Girshick, J. Donahue, T. Darrell and J. Malik. *Rich feature hierarchies for accurate object detection and semantic segmentation*. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 580–587, 2014. (Cited on page 39)

[Girshick 2015] R. Girshick. *Fast r-cnn*. In Proceedings of the IEEE international conference on computer vision, pages 1440–1448, 2015. (Cited on page 39)

[Gkioxari 2013] G. Gkioxari, P. Arbeláez, L. Bourdev and J. Malik. *Articulated pose estimation using discriminative armlet classifiers*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3342–3349, 2013. (Cited on pages xiv and 37)

[Gochoo 2020] M. Gochoo, T.-H. Tan, F. Alnajjar, J.-W. Hsieh and P.-Y. Chen. *Lownet: Privacy Preserved Ultra-Low Resolution Posture Image Classification*. In 2020 IEEE International Conference on Image Processing (ICIP), pages 663–667. IEEE, 2020. (Cited on page 23)

[Goldenberg 2017] M. G. Goldenberg, J. Jung and T. P. Grantcharov. *Using data to enhance performance and improve quality and safety in surgery*. JAMA surgery, vol. 152, no. 10, pages 972–973, 2017. (Cited on page 9)

[Goodfellow 2014] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio. *Generative adversarial nets*. In Proceedings of the 27th International Conference on Neural Information Processing Systems-Volume 2, pages 2672–2680, 2014. (Cited on page 20)

[Goyal 2017] P. Goyal, P. Dollár, R. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Y. Jia and K. He. *Accurate, large minibatch sgd: Training imagenet in 1 hour*. arXiv preprint arXiv:1706.02677, 2017. (Cited on page 82)

[Grill 2020] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. D. Guo, M. G. Azar*et al. Bootstrap your own latent: A new approach to self-supervised learning*. arXiv preprint arXiv:2006.07733, 2020. (Cited on pages 12 and 21)

[Hansen 2019] L. Hansen, M. Siebert, J. Diesel and M. P. Heinrich. *Fusing information from multiple 2D depth cameras for 3D human pose estimation in the operating room*. International journal of computer assisted radiology and surgery, vol. 14, no. 11, pages 1871–1879, 2019. (Cited on page 26)

[Haque 2016] A. Haque, B. Peng, Z. Luo, A. Alahi, S. Yeung and L. Fei-Fei. *Towards viewpoint invariant 3d human pose estimation*. In ECCV, pages 160–177. Springer, 2016. (Cited on page 49)

[Haque 2017] A. Haque, M. Guo, A. Alahi, S. Yeung, Z. Luo, A. Rege, J. Jopling, L. Downing, W. Beninati, A. Singh*et al. Towards Vision-Based Smart Hospitals: A System for Tracking and Monitoring Hand Hygiene Compliance.* In Proceedings of Machine Learning for Healthcare, volume 68, 2017. (Cited on page 23)

[Haris 2018] M. Haris, G. Shakhnarovich and N. Ukita. *Task-driven super resolution: Object detection in low-resolution images.* arXiv preprint arXiv:1803.11316, 2018. (Cited on page 22)

[He 2016] K. He, X. Zhang, S. Ren and J. Sun. *Deep residual learning for image recognition.* In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016. (Cited on pages 5, 52, 73, 81, and 87)

[He 2017] K. He, G. Gkioxari, P. Dollár and R. Girshick. *Mask R-CNN.* In Proceedings of the IEEE international conference on computer vision, pages 2961–2969, 2017. (Cited on pages xiii, xxii, xxiii, 5, 6, 13, 14, 25, 26, 27, 38, 39, 40, 49, 50, 52, 53, 60, 71, 73, 74, 75, 82, 83, 87, 90, 92, 134, 136, and 137)

[He 2020] K. He, H. Fan, Y. Wu, S. Xie and R. Girshick. *Momentum contrast for unsupervised visual representation learning.* In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9729–9738, 2020. (Cited on pages xxii, 12, 21, 75, 81, and 83)

[Helmreich 2000] R. L. Helmreich. *On error management: lessons from aviation.* Bmj, vol. 320, no. 7237, pages 781–785, 2000. (Cited on page 9)

[Hesse 2018a] N. Hesse, C. Bodensteiner, M. Arens, U. G. Hofmann, R. Weinberger and A. Sebastian Schroeder. *Computer vision for medical infant motion analysis: State of the art and rgb-d data set.* In Proceedings of the European Conference on Computer Vision (ECCV) Workshops, pages 0–0, 2018. (Cited on page 100)

[Hesse 2018b] N. Hesse, S. Pujades, J. Romero, M. J. Black, C. Bodensteiner, M. Arens, U. G. Hofmann, U. Tacke, M. Hadders-Algra, R. Weinberger*et al. Learning an infant body model from RGB-D data for accurate full body motion analysis.* In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 792–800. Springer, 2018. (Cited on page 100)

[Hinton 2015] G. Hinton, O. Vinyals and J. Dean. *Distilling the knowledge in a neural network.* arXiv preprint arXiv:1503.02531, 2015. (Cited on pages 14, 27, 60, 80, and 135)

[Hoffman 2016] J. Hoffman, D. Wang, F. Yu and T. Darrell. *Fcns in the wild: Pixel-level adversarial and constraint-based adaptation.* arXiv preprint arXiv:1612.02649, 2016. (Cited on page 20)

[Hsu 2020] C.-C. Hsu, Y.-H. Tsai, Y.-Y. Lin and M.-H. Yang. *Every pixel matters: Center-aware feature alignment for domain adaptive object detector.* In European

Conference on Computer Vision, pages 733–748. Springer, 2020. (Cited on page 20)

[Hu 2017] P. Hu and D. Ramanan. *Finding tiny faces*. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 951–959, 2017. (Cited on pages 117, 119, 122, 123, and 124)

[Inoue 2018] N. Inoue, R. Furuta, T. Yamasaki and K. Aizawa. *Cross-domain weakly-supervised object detection through progressive domain adaptation*. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 5001–5009, 2018. (Cited on page 21)

[Insafutdinov 2016] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka and B. Schiele. *Deepercut: A deeper, stronger, and faster multi-person pose estimation model*. In European Conference on Computer Vision, pages 34–50. Springer, 2016. (Cited on page 115)

[Ionescu 2013] C. Ionescu, D. Papava, V. Olaru and C. Sminchisescu. *Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments*. IEEE transactions on pattern analysis and machine intelligence, vol. 36, no. 7, pages 1325–1339, 2013. (Cited on pages 24, 27, 32, 40, 62, 108, and 142)

[Issenhuth 2019] T. Issenhuth, V. Srivastav, A. Gangi and N. Padoy. *Face detection in the operating room: comparison of state-of-the-art methods and a self-supervised approach*. International Journal of Computer Assisted Radiology and Surgery, vol. 14, no. 117, page 10, Jun 2019. Presented in the 10th International Conference on Information Processing in Computer-Assisted Interventions (IPCAI 2019). (Cited on pages 15 and 114)

[Jackson 1915] C. Jackson. Peroral endoscopy and laryngeal surgery. Laryngoscope Company, 1915. (Cited on pages xiii and 7)

[Jaffray 2005] B. Jaffray. *Minimally invasive surgery*. Archives of disease in childhood, vol. 90, no. 5, pages 537–542, 2005. (Cited on page 9)

[Jiang 2017] H. Jiang and E. Learned-Miller. *Face detection with the faster R-CNN*. In 2017 12th IEEE international conference on automatic face & gesture recognition (FG 2017), pages 650–657. IEEE, 2017. (Cited on pages 114, 115, 117, and 122)

[Jin 2020] S. Jin, L. Xu, J. Xu, C. Wang, W. Liu, C. Qian, W. Ouyang and P. Luo. *Whole-body human pose estimation in the wild*. In European Conference on Computer Vision, pages 196–214. Springer, 2020. (Cited on page 99)

[Johansson 1973] G. Johansson. *Visual perception of biological motion and a model for its analysis*. Perception & psychophysics, vol. 14, no. 2, pages 201–211, 1973. (Cited on pages 23 and 24)

[Joo 2015] H. Joo, H. Liu, L. Tan, L. Gui, B. Nabbe, I. Matthews, T. Kanade, S. Nobuhara and Y. Sheikh. *Panoptic studio: A massively multiview system for social motion capture.* In Proceedings of the IEEE International Conference on Computer Vision, pages 3334–3342, 2015. (Cited on pages 107 and 140)

[Kadkhodamohammadi 2014] A. Kadkhodamohammadi, A. Gangi, M. de Mathelin and N. Padoy. *Temporally consistent 3D pose estimation in the interventional room using discrete MRF optimization over RGB-D sequences.* In International Conference on Information Processing in Computer-Assisted Interventions, pages 168–177. Springer, 2014. (Cited on page 25)

[Kadkhodamohammadi 2015] A. Kadkhodamohammadi, A. Gangi, M. de Mathelin and N. Padoy. *Pictorial structures on RGB-D images for human pose estimation in the operating room.* In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 363–370. Springer, 2015. (Cited on page 25)

[Kadkhodamohammadi 2017a] A. Kadkhodamohammadi, A. Gangi, M. de Mathelin and N. Padoy. *Articulated clinician detection using 3D pictorial structures on RGB-D data.* Medical image analysis, vol. 35, pages 215–224, 2017. (Cited on page 26)

[Kadkhodamohammadi 2017b] A. Kadkhodamohammadi, A. Gangi, M. de Mathelin and N. Padoy. *Articulated clinician detection using 3D pictorial structures on RGB-D data.* Medical image analysis, vol. 35, pages 215–224, 2017. (Cited on page 115)

[Kadkhodamohammadi 2017c] A. Kadkhodamohammadi, A. Gangi, M. de Mathelin and N. Padoy. *A multi-view RGB-D approach for human pose estimation in operating rooms.* In 2017 IEEE winter conference on applications of computer vision (WACV), pages 363–372. IEEE, 2017. (Cited on pages 26, 115, and 130)

[Kanazawa 2018] A. Kanazawa, M. J. Black, D. W. Jacobs and J. Malik. *End-to-end recovery of human shape and pose.* In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 7122–7131, 2018. (Cited on pages 108 and 142)

[Kannan 2019] S. Kannan, G. Yengera, D. Mutter, J. Marescaux and N. Padoy. *Future-state predicting LSTM for early surgery type recognition.* IEEE transactions on medical imaging, vol. 39, no. 3, pages 556–566, 2019. (Cited on page 9)

[Khodabandeh 2019] M. Khodabandeh, A. Vahdat, M. Ranjbar and W. G. Macready. *A robust learning approach to domain adaptive object detection.* In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 480–490, 2019. (Cited on page 21)

[Khurana 2021] S. Khurana, N. Moritz, T. Hori and J. Le Roux. *Unsupervised domain adaptation for speech recognition via uncertainty driven self-training.* In

ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 6553–6557. IEEE, 2021. (Cited on page 19)

[Kim 2019] S. Kim, J. Choi, T. Kim and C. Kim. *Self-training and adversarial background regularization for unsupervised domain adaptive one-stage object detection.* In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 6092–6101, 2019. (Cited on page 21)

[Kirillov 2017] A. Kirillov, E. Levinkov, B. Andres, B. Savchynskyy and C. Rother. *Instancecut: from edges to instances with multicut.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 5008–5017, 2017. (Cited on page 26)

[Kocabas 2020] M. Kocabas, N. Athanasiou and M. J. Black. *Vibe: Video inference for human body pose and shape estimation.* In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5253–5263, 2020. (Cited on pages 108 and 142)

[Konkle 2021] T. Konkle and G. A. Alvarez. *Beyond category-supervision: instance-level contrastive learning models predict human visual system responses to objects.* bioRxiv, 2021. (Cited on page 12)

[Krebs 2021] A. Krebs, C. Rolland, J. Verde and N. Padoy. *Pose optimization of X-ray protective shields for staff radiation exposure minimization.* International Conference on Information Processing in Computer-Assisted Interventions, 2021. (Cited on pages 10 and 98)

[Kreiss 2019] S. Kreiss, L. Bertoni and A. Alahi. *Pifpaf: Composite fields for human pose estimation.* In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 11977–11986, 2019. (Cited on page 24)

[Krizhevsky 2012] A. Krizhevsky, I. Sutskever and G. E. Hinton. *Imagenet classification with deep convolutional neural networks.* Advances in neural information processing systems, vol. 25, pages 1097–1105, 2012. (Cited on page 5)

[Kurzweil 2005] R. Kurzweil. The singularity is near: When humans transcend biology. Penguin, 2005. (Cited on page 4)

[Ladikos 2010] A. Ladikos, C. Cagniart, R. Ghotbi, M. Reiser and N. Navab. *Estimating radiation exposure in interventional environments.* In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 237–244. Springer, 2010. (Cited on page 10)

[Laine 2016] S. Laine and T. Aila. *Temporal ensembling for semi-supervised learning.* arXiv preprint arXiv:1610.02242, 2016. (Cited on page 116)

[LeCun 1989] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard and L. D. Jackel. *Backpropagation applied to handwritten zip code recognition.* Neural computation, vol. 1, no. 4, pages 541–551, 1989. (Cited on page 5)

[Lee 2013] D.-H. Lee*et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks.* In Workshop on challenges in representation learning, ICML, volume 3, 2013. (Cited on page 79)

[Lee 2020] Y. Lee and J. Park. *Centermask: Real-time anchor-free instance segmentation.* In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 13906–13915, 2020. (Cited on page 26)

[Li 2017] J. Li, X. Liang, Y. Wei, T. Xu, J. Feng and S. Yan. *Perceptual generative adversarial networks for small object detection.* In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1222–1230, 2017. (Cited on page 22)

[Li 2019] Y. Li, L. Yuan and N. Vasconcelos. *Bidirectional learning for domain adaptation of semantic segmentation.* In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6936–6945, 2019. (Cited on page 20)

[Li 2020a] H. Li, T. Loehr, A. Sekuboyina, J. Zhang, B. Wiestler and B. Menze. *Domain Adaptive Medical Image Segmentation via Adversarial Learning of Disease-Specific Spatial Patterns.* arXiv preprint arXiv:2001.09313, 2020. (Cited on page 22)

[Li 2020b] Z. Li, A. Shaban, J.-G. Simard, D. Rabindran, S. DiMaio and O. Mohareri. *A Robotic 3D Perception System for Operating Room Environment Awareness.* arXiv preprint arXiv:2003.09487, 2020. (Cited on page 26)

[Li 2021] X. Li, M. Jiang, X. Zhang, M. Kamp and Q. Dou. *Fedbn: Federated learning on non-iid features via local batch normalization.* arXiv preprint arXiv:2102.07623, 2021. (Cited on page 21)

[Liang 2017] X. Liang, L. Lin, Y. Wei, X. Shen, J. Yang and S. Yan. *Proposal-free network for instance-level object segmentation.* IEEE transactions on pattern analysis and machine intelligence, vol. 40, no. 12, pages 2978–2991, 2017. (Cited on page 26)

[Liang 2019] J. Liang, R. He, Z. Sun and T. Tan. *Exploring uncertainty in pseudo-label guided unsupervised domain adaptation.* Pattern Recognition, vol. 96, page 106996, 2019. (Cited on page 21)

[Lin 2014] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár and C. L. Zitnick. *Microsoft coco: Common objects in context.* In European conference on computer vision, pages 740–755. Springer, 2014. (Cited on pages xiv, xviii, 12, 32, 37, 41, 48, 49, 62, 70, 80, 115, 117, 119, and 130)

[Lin 2017a] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan and S. Belongie. *Feature pyramid networks for object detection.* In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2117–2125, 2017. (Cited on pages 52 and 73)

[Lin 2017b] T.-Y. Lin, P. Goyal, R. Girshick, K. He and P. Dollár. *Focal loss for dense object detection.* In Proceedings of the IEEE international conference on computer vision, pages 2980–2988, 2017. (Cited on page 38)

[Liu 2010] F. Liu, D. Wang, B. Li and Y. Liu. *Improving blog polarity classification via topic analysis and adaptive methods.* In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pages 309–312, 2010. (Cited on page 20)

[Liu 2016] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu and A. C. Berg. *Ssd: Single shot multibox detector.* In European conference on computer vision, pages 21–37. Springer, 2016. (Cited on page 117)

[Liu 2017] S. Liu, J. Jia, S. Fidler and R. Urtasun. *Sgn: Sequential grouping networks for instance segmentation.* In Proceedings of the IEEE International Conference on Computer Vision, pages 3496–3504, 2017. (Cited on page 26)

[Liu 2021a] D. Liu, Q. Li, T. Jiang, Y. Wang, R. Miao, F. Shan and Z. Li. *Towards Unified Surgical Skill Assessment.* In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9522–9531, 2021. (Cited on page 99)

[Liu 2021b] Y.-C. Liu, C.-Y. Ma, Z. He, C.-W. Kuo, K. Chen, P. Zhang, B. Wu, Z. Kira and P. Vajda. *Unbiased teacher for semi-supervised object detection.* arXiv preprint arXiv:2102.09480, 2021. (Cited on pages xxiii, 21, 71, 74, 82, 91, and 92)

[Loper 2015] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll and M. J. Black. *SMPL: A skinned multi-person linear model.* ACM transactions on graphics (TOG), vol. 34, no. 6, pages 1–16, 2015. (Cited on pages 108 and 142)

[Loy Rodas 2018] N. Loy Rodas. *Context-aware radiation protection for the hybrid operating room.* PhD thesis, Strasbourg, 2018. (Cited on pages xvii, 97, and 98)

[Mahapatra 2019] D. Mahapatra, B. Bozorgtabar and R. Garnavi. *Image super-resolution using progressive generative adversarial networks for medical image analysis.* Computerized Medical Imaging and Graphics, vol. 71, pages 30–39, 2019. (Cited on page 23)

[Maier-Hein 2017] L. Maier-Hein, S. Vedula, S. Speidel, N. Navab, R. Kikinis, A. Park, M. Eisenmann, H. Feussner, G. Forestier, S. Giannarou*et al. Surgical data science: enabling next-generation surgery.* Nature Biomedical Engineering, vol. 1, pages 691–696, 2017. (Cited on page 115)

[Makary 2016] M. A. Makary and M. Daniel. *Medical error—the third leading cause of death in the US.* Bmj, vol. 353, 2016. (Cited on page 10)

[Mao 2021] W. Mao, Z. Tian, X. Wang and C. Shen. *FCPose: Fully Convolutional Multi-Person Pose Estimation with Dynamic Instance-Aware Convolutions.* In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9034–9043, 2021. (Cited on page 24)

[Martinez 2017] J. Martinez, R. Hossain, J. Romero and J. J. Little. *A simple yet effective baseline for 3d human pose estimation.* In Proceedings of the IEEE International Conference on Computer Vision, pages 2640–2649, 2017. (Cited on pages xxi, 25, 27, 40, 43, 45, and 62)

[Mascagni 2021a] P. Mascagni, D. Alapatt, A. Garcia, N. Okamoto, A. Vardazaryan, G. Costamagna, B. Dallemagne and N. Padoy. *Surgical data science for safe cholecystectomy: a protocol for segmentation of hepatocystic anatomy and assessment of the critical view of safety.* arXiv preprint arXiv:2106.10916, 2021. (Cited on page 9)

[Mascagni 2021b] P. Mascagni, D. Alapatt, T. Urade, A. Vardazaryan, D. Mutter, J. Marescaux, G. Costamagna, B. Dallemagne and N. Padoy. *A computer vision platform to automatically locate critical events in surgical videos: documenting safety in laparoscopic cholecystectomy.* Annals of Surgery, vol. 274, no. 1, pages e93–e95, 2021. (Cited on page 9)

[Mascagni 2021c] P. Mascagni and N. Padoy. *OR black box and surgical control tower: Recording and streaming data and analytics to improve surgical care.* Journal of Visceral Surgery, 2021. (Cited on pages 3 and 9)

[Mascagni 2021d] P. Mascagni, M. R. Rodriguez-Luna, T. Urade, E. Felli, P. Pessaux, D. Mutter, J. Marescaux, G. Costamagna, B. Dallemagne and N. Padoy. *Intraoperative time out to promote the implementation of the critical view of safety in laparoscopic cholecystectomy: a video-based assessment of 343 procedures.* arXiv preprint arXiv:2104.02338, 2021. (Cited on page 9)

[Mascagni 2021e] P. Mascagni, A. Vardazaryan, D. Alapatt, T. Urade, T. Emre, C. Fiorillo, P. Pessaux, D. Mutter, J. Marescaux, G. Costamagna*et al. Artificial intelligence for surgical safety: automatic assessment of the critical view of safety in laparoscopic cholecystectomy using deep learning.* Annals of Surgery, 2021. (Cited on page 9)

[McInnes 2018] L. McInnes, J. Healy and J. Melville. *Umap: Uniform manifold approximation and projection for dimension reduction.* arXiv preprint arXiv:1802.03426, 2018. (Cited on pages xiv, xviii, 12, and 130)

[McMahan 2017] B. McMahan, E. Moore, D. Ramage, S. Hampson and B. A. y Arcas. *Communication-efficient learning of deep networks from decentralized data.* In Artificial intelligence and statistics, pages 1273–1282. PMLR, 2017. (Cited on page 22)

[McNally 2020] W. McNally, K. Vats, A. Wong and J. McPhee. *EvoPose2D: Pushing the Boundaries of 2D Human Pose Estimation using Neuroevolution.* arXiv preprint arXiv:2011.08446, 2020. (Cited on page 24)

[Mehta 2017] D. Mehta, O. Sotnychenko, F. Mueller, W. Xu, S. Sridhar, G. Pons-Moll and C. Theobalt. *Single-shot multi-person 3d body pose estimation from monocular rgb input.* arXiv preprint arXiv:1712.03453, 2017. (Cited on page 25)

[Meißner 2014] C. Meißner, J. Meixensberger, A. Pretschner and T. Neumuth. *Sensor-based surgical activity recognition in unconstrained environments.* Minimally Invasive Therapy & Allied Technologies, vol. 23, no. 4, pages 198–205, 2014. (Cited on page 9)

[Meng 2019] Z. Meng, J. Li, Y. Gaur and Y. Gong. *Domain adaptation via teacher-student learning for end-to-end speech recognition.* In 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pages 268–275. IEEE, 2019. (Cited on page 19)

[Micikevicius 2017] P. Micikevicius, S. Narang, J. Alben, G. Diamos, E. Elsen, D. Garcia, B. Ginsburg, M. Houston, O. Kuchaiev, G. Venkatesh*et al. Mixed precision training.* arXiv preprint arXiv:1710.03740, 2017. (Cited on page 82)

[Misra 2020] I. Misra and L. v. d. Maaten. *Self-supervised learning of pretext-invariant representations.* In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6707–6717, 2020. (Cited on page 75)

[Moravec 1998] H. Moravec. *When will computer hardware match the human brain.* Journal of evolution and technology, vol. 1, no. 1, page 10, 1998. (Cited on page 4)

[Moreno-Noguer 2017] F. Moreno-Noguer. *3d human pose estimation from a single image via distance matrix regression.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2823–2832, 2017. (Cited on page 25)

[Najibi 2017] M. Najibi, P. Samangouei, R. Chellappa and L. S. Davis. *Ssh: Single stage headless face detector.* In Proceedings of the IEEE international conference on computer vision, pages 4875–4884, 2017. (Cited on pages xvii, 114, 116, 117, 118, 119, 122, 123, and 124)

[Nara 2010] A. Nara, K. Izumi, H. Iseki, T. Suzuki, K. Nambu and Y. Sakurai. *Surgical workflow monitoring based on trajectory data mining.* In JSAI International

Symposium on Artificial Intelligence, pages 283–291. Springer, 2010. (Cited on page 9)

[Neumann 2018] L. Neumann and A. Vedaldi. *Tiny people pose*. In Asian Conference on Computer Vision, pages 558–574. Springer, 2018. (Cited on page 22)

[Newell 2017] A. Newell, Z. Huang and J. Deng. *Associative embedding: End-to-end learning for joint detection and grouping*. Advances in Neural Information Processing Systems, vol. 2017, pages 2278–2288, 2017. (Cited on page 24)

[Nieto-Rodríguez 2015] A. Nieto-Rodríguez, M. Mucientes and V. M. Brea. *System for Medical Mask Detection in the Operating Room Through Facial Attributes*. In Iberian Conference on Pattern Recognition and Image Analysis, pages 138–145. Springer, 2015. (Cited on page 115)

[Norton 2000] D. F. Norton and M. J. Norton. *A Treatise of Human Nature: Being an Attempt to Introduce the Experimental Method of Reasoning Into Moral Subjects*. 2000. (Cited on page 12)

[Nowak 2020] M. Nowak, P. Carbonez, M. Krauss, F. Verdun and J. Damet. *Characterisation and mapping of scattered radiation fields in interventional radiology theatres*. Scientific Reports, vol. 10, no. 1, pages 1–9, 2020. (Cited on page 10)

[Nwoye 2019] C. I. Nwoye, D. Mutter, J. Marescaux and N. Padoy. *Weakly supervised convolutional LSTM approach for tool tracking in laparoscopic videos*. International journal of computer assisted radiology and surgery, vol. 14, no. 6, pages 1059–1067, 2019. (Cited on page 9)

[Nwoye 2020] C. I. Nwoye, C. Gonzalez, T. Yu, P. Mascagni, D. Mutter, J. Marescaux and N. Padoy. *Recognition of instrument-tissue interactions in endoscopic videos via action triplets*. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 364–374. Springer, 2020. (Cited on page 9)

[Orbes-Arteainst 2019] M. Orbes-Arteainst, J. Cardoso, L. Sørensen, C. Igel, S. Ourselin, M. Modat, M. Nielsen and A. Pai. *Knowledge distillation for semi-supervised domain adaptation*. In OR 2.0 Context-Aware Operating Theaters and Machine Learning in Clinical Neuroimaging, pages 68–76. Springer, 2019. (Cited on page 22)

[Ouyang 2019] C. Ouyang, K. Kamnitsas, C. Biffi, J. Duan and D. Rueckert. *Data efficient unsupervised domain adaptation for cross-modality image segmentation*. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 669–677. Springer, 2019. (Cited on page 22)

[Oza 2021] P. Oza, V. A. Sindagi, V. VS and V. M. Patel. *Unsupervised Domain Adaption of Object Detectors: A Survey.* arXiv preprint arXiv:2105.13502, 2021. (Cited on pages 22 and 71)

[Padoy 2009] N. Padoy, D. Mateus, D. Weinland, M.-O. Berger and N. Navab. *Workflow monitoring based on 3d motion features.* In 2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops, pages 585–592. IEEE, 2009. (Cited on page 9)

[Padoy 2012] N. Padoy, T. Blum, S.-A. Ahmadi, H. Feussner, M.-O. Berger and N. Navab. *Statistical modeling and recognition of surgical workflow.* Medical image analysis, vol. 16, no. 3, pages 632–641, 2012. (Cited on page 9)

[Padoy 2018] N. Padoy and B. Villani. *Vers une tour de contrôle des blocs opératoires? Santé Intell. Artif.* 2018. (Cited on page 9)

[Papandreou 2018] G. Papandreou, T. Zhu, L.-C. Chen, S. Gidaris, J. Tompson and K. Murphy. *Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model.* In Proceedings of the European Conference on Computer Vision (ECCV), pages 269–286, 2018. (Cited on page 26)

[Patel 2015] V. M. Patel, R. Gopalan, R. Li and R. Chellappa. *Visual domain adaptation: A survey of recent advances.* IEEE signal processing magazine, vol. 32, no. 3, pages 53–69, 2015. (Cited on page 22)

[Peng 2018] C. Peng, T. Xiao, Z. Li, Y. Jiang, X. Zhang, K. Jia, G. Yu and J. Sun. *Megdet: A large mini-batch object detector.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 6181–6189, 2018. (Cited on pages xxii, 75, 81, and 83)

[Powles 2017] J. Powles and H. Hodson. *Google DeepMind and healthcare in an age of algorithms.* Health and technology, vol. 7, no. 4, pages 351–367, 2017. (Cited on pages 8 and 22)

[Prechtl 1990] H. F. Prechtl. *Qualitative changes of spontaneous movements in fetus and preterm infant are a marker of neurological dysfunction.* Early human development, 1990. (Cited on page 100)

[Radosavovic 2018] I. Radosavovic, P. Dollár, R. Girshick, G. Gkioxari and K. He. *Data distillation: Towards omni-supervised learning.* In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 4119–4128, 2018. (Cited on pages 14, 27, 60, 61, 79, 80, 116, and 135)

[Ramesh 2021] S. Ramesh, D. Dall'Alba, C. Gonzalez, T. Yu, P. Mascagni, D. Mutter, J. Marescaux, P. Fiorini and N. Padoy. *Multi-task temporal convolutional networks*

for joint recognition of surgical phases and steps in gastric bypass procedures. International Journal of Computer Assisted Radiology and Surgery, pages 1–9, 2021. (Cited on page 9)

[Reddy 2021] N. D. Reddy, L. Guigues, L. Pishchulin, J. Eledath and S. G. Narasimhan. *TesseTrack: End-to-End Learnable Multi-Person Articulated 3D Pose Tracking.* In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 15190–15200, 2021. (Cited on pages 107 and 140)

[Reiley 2011] C. E. Reiley, H. C. Lin, D. D. Yuh and G. D. Hager. *Review of methods for objective surgical skill evaluation.* Surgical endoscopy, vol. 25, no. 2, pages 356–366, 2011. (Cited on page 10)

[Ren 2015] S. Ren, K. He, R. Girshick and J. Sun. *Faster r-cnn: Towards real-time object detection with region proposal networks.* Advances in neural information processing systems, vol. 28, pages 91–99, 2015. (Cited on pages 38, 39, 75, 115, 117, 118, and 122)

[Rodas 2015] N. L. Rodas and N. Padoy. *Seeing is believing: increasing intraoperative awareness to scattered radiation in interventional procedures by combining augmented reality, Monte Carlo simulations and wireless dosimeters.* International journal of computer assisted radiology and surgery, vol. 10, no. 8, pages 1181–1191, 2015. (Cited on page 33)

[Rodas 2016] N. L. Rodas, F. Barrera and N. Padoy. *See it with your own eyes: Markerless mobile augmented reality for radiation awareness in the hybrid room.* IEEE Transactions on Biomedical Engineering, vol. 64, no. 2, pages 429–440, 2016. (Cited on pages 10 and 98)

[Rodas 2017] N. L. Rodas, F. Barrera and N. Padoy. *See it with your own eyes: markerless mobile augmented reality for radiation awareness in the hybrid room.* IEEE Transactions on Biomedical Engineering, vol. 64, no. 2, pages 429–440, 2017. (Cited on page 130)

[Rogez 2017] G. Rogez, P. Weinzaepfel and C. Schmid. *Lcr-net: Localization-classification-regression for human pose.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3433–3441, 2017. (Cited on page 25)

[Roguin 2013] A. Roguin, J. Goldstein, O. Bar and J. A. Goldstein. *Brain and neck tumors among physicians performing interventional procedures.* The American journal of cardiology, vol. 111, no. 9, pages 1368–1372, 2013. (Cited on page 10)

[Ross 2017] T.-Y. Ross and G. Dollár. *Focal loss for dense object detection.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2980–2988, 2017. (Cited on pages 53 and 74)

[RoyChowdhury 2019] A. RoyChowdhury, P. Chakrabarty, A. Singh, S. Jin, H. Jiang, L. Cao and E. Learned-Miller. *Automatic adaptation of object detectors to new domains using self-training.* In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 780–790, 2019. (Cited on page 21)

[Ruggero Ronchi 2017] M. Ruggero Ronchi and P. Perona. *Benchmarking and error diagnosis in multi-instance pose estimation.* In Proceedings of the IEEE international conference on computer vision, pages 369–378, 2017. (Cited on pages xvi, 84, and 89)

[Rumelhart 1986] D. E. Rumelhart, G. E. Hinton and R. J. Williams. *Learning representations by back-propagating errors.* nature, vol. 323, no. 6088, pages 533–536, 1986. (Cited on page 5)

[Russakovsky 2015] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein*et al. Imagenet large scale visual recognition challenge.* International journal of computer vision, vol. 115, no. 3, pages 211–252, 2015. (Cited on pages 4 and 5)

[Rutkow 2000] I. M. Rutkow. *Trephination.* Archives of Surgery, vol. 135, no. 9, pages 1119–1119, 2000. (Cited on page 7)

[Ryoo 2017] M. S. Ryoo, B. Rothrock, C. Fleming and H. J. Yang. *Privacy-preserving human activity recognition from extreme low resolution.* In Thirty-First AAAI Conference on Artificial Intelligence, 2017. (Cited on page 22)

[Saito 2019] K. Saito, Y. Ushiku, T. Harada and K. Saenko. *Strong-weak distribution alignment for adaptive object detection.* In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6956–6965, 2019. (Cited on page 20)

[Sajjadi 2016] M. Sajjadi, M. Javanmardi and T. Tasdizen. *Regularization with stochastic transformations and perturbations for deep semi-supervised learning.* Advances in neural information processing systems, vol. 29, pages 1163–1171, 2016. (Cited on page 72)

[Saxe 2013] A. M. Saxe, J. L. McClelland and S. Ganguli. *Exact solutions to the nonlinear dynamics of learning in deep linear neural networks.* arXiv:1312.6120, 2013. (Cited on page 53)

[Sestini 2021] L. Sestini, B. Rosa, E. De Momi, G. Ferrigno and N. Padoy. *A Kinematic Bottleneck Approach For Pose Regression of Flexible Surgical Instruments directly from Images.* IEEE Robotics and Automation Letters, vol. 6, no. 2, pages 2938–2945, 2021. (Cited on page 9)

[Sheller 2018] M. J. Sheller, G. A. Reina, B. Edwards, J. Martin and S. Bakas. *Multi-institutional deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation.* In International MICCAI Brainlesion Workshop, pages 92–104. Springer, 2018. (Cited on page 22)

[Shermeyer 2019] J. Shermeyer and A. Van Etten. *The effects of super-resolution on object detection performance in satellite imagery.* In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pages 0–0, 2019. (Cited on page 22)

[Shi 2016] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert and Z. Wang. *Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network.* In CVPR, pages 1874–1883, 2016. (Cited on page 52)

[Shotton 2013] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook and R. Moore. *Real-time human pose recognition in parts from single depth images.* Communications of the ACM, vol. 56, no. 1, pages 116–124, 2013. (Cited on page 49)

[Sigal 2010] L. Sigal, A. O. Balan and M. J. Black. *Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion.* International journal of computer vision, vol. 87, no. 1-2, page 4, 2010. (Cited on pages 24 and 32)

[Sindagi 2020] V. A. Sindagi, P. Oza, R. Yasarla and V. M. Patel. *Prior-based domain adaptive object detection for hazy and rainy conditions.* In European Conference on Computer Vision, pages 763–780. Springer, 2020. (Cited on page 20)

[Soenens 2021] G. Soenens, B. Doyen, P. Vlerick, F. Vermassen, T. Grantcharov and I. Van Herzeele. *Assessment of Endovascular Team Performances Using a Comprehensive Data Capture Platform in the Hybrid Room: A Pilot Study.* European Journal of Vascular and Endovascular Surgery, vol. 61, no. 6, pages 1028–1029, 2021. (Cited on page 11)

[Sohn 2020a] K. Sohn, D. Berthelot, C.-L. Li, Z. Zhang, N. Carlini, E. D. Cubuk, A. Kurakin, H. Zhang and C. Raffel. *Fixmatch: Simplifying semi-supervised learning with consistency and confidence.* arXiv preprint arXiv:2001.07685, 2020. (Cited on pages 21, 71, and 72)

[Sohn 2020b] K. Sohn, Z. Zhang, C.-L. Li, H. Zhang, C.-Y. Lee and T. Pfister. *A simple semi-supervised learning framework for object detection.* arXiv preprint arXiv:2005.04757, 2020. (Cited on page 79)

[Srivastav 2018] V. Srivastav, T. Issenhuth, K. Abdolrahim, M. de Mathelin, A. Gangi and N. Padoy. *MVOR: A Multi-view RGB-D Operating Room Dataset for 2D*

and 3D Human Pose Estimation. In MICCAI-LABELS workshop, 2018. (Cited on pages xiv, xviii, xxiii, 12, 15, 116, 119, 122, and 130)

[Srivastav 2019] V. Srivastav, A. Gangi and N. Padoy. *Human Pose Estimation on Privacy-Preserving Low-Resolution Depth Images.* In MICCAI, pages 583–591. Springer, 2019. (Cited on page 15)

[Srivastav 2020] V. Srivastav, A. Gangi and N. Padoy. *Self-supervision on Unlabelled OR Data for Multi-person 2D/3D Human Pose Estimation.* In International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2020. (Cited on page 15)

[Sun 2019] K. Sun, B. Xiao, D. Liu and J. Wang. *Deep High-Resolution Representation Learning for Human Pose Estimation.* In Proceedings of the IEEE conference on computer vision and pattern recognition, 2019. (Cited on pages 24, 27, 39, 40, 62, and 80)

[Sun 2020] Y. Sun, Q. Bao, W. Liu, Y. Fu, M. J. Black and T. Mei. *CenterHMR: Multi-Person Center-based Human Mesh Recovery.* 2020. (Cited on pages 108 and 142)

[Svoboda 2005] T. Svoboda, D. Martinec and T. Pajdla. *A convenient multicamera self-calibration for virtual environments.* Presence: Teleoperators & virtual environments, vol. 14, no. 4, pages 407–422, 2005. (Cited on page 33)

[Symons 2017] T. Symons and T. Bass. *Me, my data and I: The future of the personal data economy.* 2017. (Cited on pages 8 and 22)

[Tan 2018] W. Tan, B. Yan and B. Bare. *Feature super-resolution: Make machine see more clearly.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3994–4002, 2018. (Cited on page 22)

[Tarvainen 2017] A. Tarvainen and H. Valpola. *Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results.* In Proceedings of the 31st International Conference on Neural Information Processing Systems, pages 1195–1204, 2017. (Cited on pages 21, 27, 72, 78, and 138)

[Tian 2019a] Z. Tian, H. Chen and C. Shen. *Directpose: Direct end-to-end multi-person pose estimation.* arXiv preprint arXiv:1911.07451, 2019. (Cited on page 24)

[Tian 2019b] Z. Tian, C. Shen, H. Chen and T. He. *Fcos: Fully convolutional one-stage object detection.* In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 9627–9636, 2019. (Cited on page 24)

[Toldo 2020] M. Toldo, A. Maracani, U. Michieli and P. Zanuttigh. *Unsupervised domain adaptation in semantic segmentation: a review.* Technologies, vol. 8, no. 2, page 35, 2020. (Cited on pages 22 and 71)

[Tran 2016] D. T. Tran, R. Sakurai and J.-H. Lee. *An improvement of surgical phase detection using latent dirichlet allocation and hidden markov model.* In Innovation in Medicine and Healthcare 2015, pages 249–261. Springer, 2016. (Cited on page 9)

[Tran 2019] L. Tran, K. Sohn, X. Yu, X. Liu and M. Chandraker. *Gotta adapt'em all: Joint pixel and feature-level domain adaptation for recognition in the wild.* In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 2672–2681, 2019. (Cited on page 20)

[Tu 2020] H. Tu, C. Wang and W. Zeng. *Voxelpose: Towards multi-camera 3d human pose estimation in wild environment.* In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16, pages 197–212. Springer, 2020. (Cited on pages 107 and 140)

[Twinanda 2015] A. P. Twinanda, E. O. Alkan, A. Gangi, M. de Mathelin and N. Padoy. *Data-driven spatio-temporal RGBD feature encoding for action recognition in operating rooms.* International journal of computer assisted radiology and surgery, vol. 10, no. 6, pages 737–747, 2015. (Cited on page 9)

[Twinanda 2016a] A. P. Twinanda, S. Shehata, D. Mutter, J. Marescaux, M. De Mathelin and N. Padoy. *Endonet: a deep architecture for recognition tasks on laparoscopic videos.* IEEE transactions on medical imaging, vol. 36, no. 1, pages 86–97, 2016. (Cited on page 9)

[Twinanda 2016b] A. P. Twinanda, S. Shehata, D. Mutter, J. Marescaux, M. de Mathelin and N. Padoy. *Multi-Stream Deep Architecture for Surgical Phase Recognition on Multi-View RGBD Videos.* In MICCAI Workshop on Modeling and Monitoring of Computer Assisted Interventions (M2CAI), 2016. (Cited on page 115)

[Van der Maaten 2008] L. Van der Maaten and G. Hinton. *Visualizing data using t-SNE.* Journal of machine learning research, vol. 9, no. 11, 2008. (Cited on pages xvii, 87, and 89)

[Vanhavere 2008] F. Vanhavere, E. Carinou, L. Donadille, M. Ginjaume, J. Jankowski, A. Rimpler and M. Sans Merce. *An overview on extremity dosimetry in medical applications.* Radiation Protection Dosimetry, vol. 129, no. 1-3, pages 350–355, 2008. (Cited on page 10)

[Viola 2001] P. Viola and M. Jones. *Rapid object detection using a boosted cascade of simple features.* In CVPR, pages I–I, 2001. (Cited on page 114)

[Vozenilek 2004] J. Vozenilek, J. S. Huff, M. Reznek and J. A. Gordon. *See one, do one, teach one: advanced technology in medical education.* Academic Emergency Medicine, vol. 11, no. 11, pages 1149–1154, 2004. (Cited on page 10)

[VS 2021] V. VS, V. Gupta, P. Oza, V. A. Sindagi and V. M. Patel. *MeGA-CDA: Memory Guided Attention for Category-Aware Unsupervised Domain Adaptive Object Detection.* In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 4516–4526, 2021. (Cited on page 20)

[Wang 2016] Z. Wang, S. Chang, Y. Yang, D. Liu and T. S. Huang. *Studying very low resolution recognition using deep networks.* In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 4792–4800, 2016. (Cited on page 22)

[Wang 2018a] M. Wang and W. Deng. *Deep visual domain adaptation: A survey.* Neurocomputing, vol. 312, pages 135–153, 2018. (Cited on page 22)

[Wang 2018b] M. Wang, X. Chen, W. Liu, C. Qian, L. Lin and L. Ma. *Drpose3d: Depth ranking in 3d human pose estimation.* arXiv preprint arXiv:1805.08973, 2018. (Cited on page 25)

[Wang 2019] X. Wang, Y. Jin, M. Long, J. Wang and M. I. Jordan. *Transferable normalization: towards improving transferability of deep neural networks.* In Proceedings of the 33rd International Conference on Neural Information Processing Systems, pages 1953–1963, 2019. (Cited on page 21)

[Wang 2020a] Q. Wang and T. Breckon. *Unsupervised domain adaptation via structured prediction based selective pseudo-labeling.* In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pages 6243–6250, 2020. (Cited on page 21)

[Wang 2020b] X. Wang, T. Kong, C. Shen, Y. Jiang and L. Li. *Solo: Segmenting objects by locations.* In European Conference on Computer Vision, pages 649–665. Springer, 2020. (Cited on page 38)

[Wang 2020c] X. Wang, R. Zhang, T. Kong, L. Li and C. Shen. *SOLOv2: Dynamic and fast instance segmentation.* arXiv preprint arXiv:2003.10152, 2020. (Cited on page 38)

[Wanzel 2002] K. R. Wanzel, E. D. Matsumoto, S. J. Hamstra and D. J. Anastakis. *Teaching technical skills: training on a simple, inexpensive, and portable model.* Plastic and reconstructive surgery, vol. 109, no. 1, pages 258–264, 2002. (Cited on page 10)

[Williams 2011] D. A. Williams. *The elephant in the room.* Music Educators Journal, vol. 98, no. 1, pages 51–57, 2011. (Cited on pages xiii and 6)

[Wu 2018] Y. Wu and K. He. *Group normalization.* In Proceedings of the European conference on computer vision (ECCV), pages 3–19, 2018. (Cited on pages xxii, 75, 81, and 83)

[Wu 2019a] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo and R. Girshick. *Detectron2*. https://github.com/facebookresearch/detectron2, 2019. (Cited on pages 65, 82, and 87)

[Wu 2019b] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo and R. Girshick. *Detectron2-keypoint-rcnn-baseline*. https://github.com/facebookresearch/detectron2/blob/master/configs/COCO-Keypoints/keypoint_rcnn_R_50_FPN_3x.yaml, 2019. (Cited on page 87)

[Wu 2019c] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo and R. Girshick. *Detectron2-maskcnn-GN-baseline*. https://github.com/facebookresearch/detectron2/blob/master/configs/Misc/mask_rcnn_R_50_FPN_3x_gn.yaml, 2019. (Cited on page 75)

[Wu 2021] Y. Wu and J. Johnson. *Rethinking" Batch" in BatchNorm*. arXiv preprint arXiv:2105.07576, 2021. (Cited on pages 21 and 75)

[Xiao 2018] B. Xiao, H. Wu and Y. Wei. *Simple baselines for human pose estimation and tracking*. In ECCV, pages 466–481, 2018. (Cited on pages 24, 39, 40, 50, and 115)

[Xie 2017] S. Xie, R. Girshick, P. Dollár, Z. Tu and K. He. *Aggregated residual transformations for deep neural networks*. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1492–1500, 2017. (Cited on pages 62 and 80)

[Xie 2019] C. Xie and A. Yuille. *Intriguing Properties of Adversarial Training at Scale*. In International Conference on Learning Representations, 2019. (Cited on page 21)

[Xie 2020] C. Xie, M. Tan, B. Gong, J. Wang, A. L. Yuille and Q. V. Le. *Adversarial examples improve image recognition*. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 819–828, 2020. (Cited on pages 21 and 75)

[Yang 2012] Y. Yang and D. Ramanan. *Articulated human detection with flexible mixtures of parts*. IEEE transactions on pattern analysis and machine intelligence, vol. 35, no. 12, pages 2878–2890, 2012. (Cited on page 41)

[Yang 2016] S. Yang, P. Luo, C.-C. Loy and X. Tang. *Wider face: A face detection benchmark*. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 5525–5533, 2016. (Cited on pages 115 and 117)

[Yang 2018] W. Yang, W. Ouyang, X. Wang, J. Ren, H. Li and X. Wang. *3d human pose estimation in the wild by adversarial learning*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 5255–5264, 2018. (Cited on page 25)

[Yeh 2019] R. Yeh, Y.-T. Hu and A. Schwing. *Chirality Nets for Human Pose Regression.* Advances in Neural Information Processing Systems, vol. 32, pages 8163–8173, 2019. (Cited on pages 71 and 77)

[Yengera 2018] G. Yengera, D. Mutter, J. Marescaux and N. Padoy. *Less is more: Surgical phase recognition with less annotations through self-supervised pre-training of CNN-LSTM networks.* arXiv preprint arXiv:1805.08569, 2018. (Cited on page 9)

[Yeung 2018] S. Yeung, N. L. Downing, L. Fei-Fei and A. Milstein. *Bedside Computer Vision-Moving Artificial Intelligence from Driver Assistance to Patient Safety.* N Engl J Med, vol. 378, no. 14, page 1271, 2018. (Cited on page 115)

[Yu 2018] T. Yu, D. Mutter, J. Marescaux and N. Padoy. *Learning from a tiny dataset of manual annotations: a teacher/student approach for surgical phase recognition.* arXiv preprint arXiv:1812.00033, 2018. (Cited on page 9)

[Zhang 2016] Z. Zhang, S. Fidler and R. Urtasun. *Instance-level segmentation for autonomous driving with deep densely connected mrfs.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 669–677, 2016. (Cited on page 26)

[Zhang 2017] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang and S. Z. Li. *S3fd: Single shot scale-invariant face detector.* In Proceedings of the IEEE international conference on computer vision, pages 192–201, 2017. (Cited on pages 114, 117, 119, 122, and 123)

[Zhang 2019a] F. Zhang, X. Zhu and M. Ye. *Fast human pose estimation.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3517–3526, 2019. (Cited on pages 14, 27, 60, 61, 80, and 135)

[Zhang 2019b] S.-H. Zhang, R. Li, X. Dong, P. Rosin, Z. Cai, X. Han, D. Yang, H. Huang and S.-M. Hu. *Pose2seg: Detection free human instance segmentation.* In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 889–898, 2019. (Cited on pages 26 and 32)

[Zhang 2020] Y. Zhang, Y. Wei, Q. Wu, P. Zhao, S. Niu, J. Huang and M. Tan. *Collaborative unsupervised domain adaptation for medical image diagnosis.* IEEE Transactions on Image Processing, vol. 29, pages 7834–7844, 2020. (Cited on page 22)

[Zhao 2020a] G. Zhao, G. Li, R. Xu and L. Lin. *Collaborative training between region proposal localization and classification for domain adaptive object detection.* In European Conference on Computer Vision, pages 86–102. Springer, 2020. (Cited on page 21)

[Zhao 2020b] S. Zhao, X. Yue, S. Zhang, B. Li, H. Zhao, B. Wu, R. Krishna, J. E. Gonzalez, A. L. Sangiovanni-Vincentelli, S. A. Seshia*et al. A review of single-source deep unsupervised visual domain adaptation.* IEEE Transactions on Neural Networks and Learning Systems, 2020. (Cited on page 22)

[Zheng 2021] Z. Zheng and Y. Yang. *Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation.* International Journal of Computer Vision, vol. 129, no. 4, pages 1106–1120, 2021. (Cited on page 21)

[Zhou 2017] X. Zhou, Q. Huang, X. Sun, X. Xue and Y. Wei. *Towards 3d human pose estimation in the wild: a weakly-supervised approach.* In Proceedings of the IEEE International Conference on Computer Vision, pages 398–407, 2017. (Cited on page 25)

[Zhou 2020] D. Zhou and Q. He. *PoSeg: Pose-aware refinement network for human instance segmentation.* IEEE Access, vol. 8, pages 15007–15016, 2020. (Cited on page 26)

[Zhu 2017] J.-Y. Zhu, T. Park, P. Isola and A. A. Efros. *Unpaired image-to-image translation using cycle-consistent adversarial networks.* In Proceedings of the IEEE international conference on computer vision, pages 2223–2232, 2017. (Cited on page 20)

[Zhuang 2020] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong and Q. He. *A comprehensive survey on transfer learning.* Proceedings of the IEEE, vol. 109, no. 1, pages 43–76, 2020. (Cited on pages 22 and 71)

[Zou 2018] Y. Zou, Z. Yu, B. Kumar and J. Wang. *Unsupervised domain adaptation for semantic segmentation via class-balanced self-training.* In Proceedings of the European conference on computer vision (ECCV), pages 289–305, 2018. (Cited on page 21)

[Zou 2019] Y. Zou, Z. Yu, X. Liu, B. Kumar and J. Wang. *Confidence regularized self-training.* In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 5982–5991, 2019. (Cited on page 21)

UNIVERSITÉ DE STRASBOURG

Vinkle Kumar Srivastav

msii ÉCOLE DOCTORALE

Unsupervised Domain Adaptation Approaches for Person Localization in the Operating Rooms

Summary

The fine-grained localization of clinicians in the operating room (OR) is a key component in designing the new OR support systems. However, the task is challenging not only because OR images contain significant visual domain differences compared to traditional vision datasets but also because data and annotations are hard to collect and generate in the OR due to privacy concerns. This thesis explores Unsupervised Domain Adaptation (UDA) methods to enable visual learning for the target domain, the OR, by working in two complementary directions. First, we study how low-resolution images with a downsampling factor as low as 12x can be used for fine-grained clinicians' localization to address privacy concerns. Second, we propose several self-supervised methods to transfer learned information from a labeled source domain to an unlabeled target domain to address the shift of visual domain and lack of annotations. These methods employ self-supervised predictions in allowing the model to learn and adapt to the unlabeled target domain. To demonstrate the effectiveness of our proposed approaches, we release the first public dataset, called the multi-view operating room (*MVOR*), generated from recordings of real clinical interventions. We obtain state-of-the-art results on the *MVOR* dataset, specifically on the privacy-preserving low-resolution OR images. We hope our proposed UDA approaches could help to scale up and deploy novel AI assistance applications for the OR environments.

Key-words: Unsupervised Domain Adaptation, Human Pose Estimation, Person Instance Segmentation, Operating Room, Low resolution Images, Semi-supervised Learning, Self-training, Depth Images.

Résumé

La localisation précise des cliniciens dans la salle d'opération est un élément clé dans la conception des nouveaux systèmes de support clinique. Cependant, la tâche est difficile non seulement parce que les images de la salle d'opération contiennent des différences visuelles significatives par rapport aux images ordinaires, mais aussi parce que les données et les annotations sont difficiles à collecter et à générer dans la salle d'opération en raison de problèmes de confidentialité. Cette thèse explore les méthodes d'adaptation de domaine non supervisées pour permettre l'apprentissage visuel pour le domaine cible, la salle d'opération, en travaillant dans deux directions complémentaires. Tout d'abord, nous étudions comment des images basse résolution avec un facteur de sous-échantillonnage allant jusqu'à 12x peuvent être utilisées pour une localisation précise des cliniciens afin de résoudre les problèmes de confidentialité. Deuxièmement, nous proposons plusieurs méthodes auto-supervisées pour transférer les informations apprises d'un domaine source étiqueté vers un domaine cible non étiqueté pour faire face au changement de domaine visuel et au manque d'annotations. Ces méthodes utilisent des prédictions auto-supervisées pour permettre au modèle d'apprendre et de s'adapter au domaine cible non étiqueté. Pour démontrer l'efficacité des approches proposées, nous publions le premier ensemble de données public, appelé Multi-View Operating Room (*MVOR)*, généré à partir d'enregistrements d'interventions cliniques réelles. Nous obtenons des résultats de pointe sur l'ensemble de données *MVOR*, en particulier sur les images de salle d'opération à basse résolution préservant la confidentialité. Nous espérons que nos approches d'adaptation de domaine non supervisées proposées pourront aider à développer et à déployer de nouvelles applications d'assistance par IA pour les salles d'opération.

Mots-clés: Adaptation de domaine non supervisée, estimation de pose humaine, segmentation d'instance de personne, salle d'opération, images basse résolution, apprentissage semi-supervisé, auto-formation, images de profondeur.