Towards Increasing Robustness Against Occlusions for Preterm Infant Pose Estimation in Videos

Roberto A. Navarro San Martin

August 24, 2022

Acknowledgements

I would like to thank my parents, Juan and Lourdes, for always supporting and helping me throughout the entire process; always reminding me that there is more to life than work and pushing me to go outside and live a little. Jussara, my partner, for keeping me grounded and being an unwavering source of support. I would also like to thank my sister, Oriana, for always being a phone call away and always reaching out to me during the most stressful and intense moments of this research. Additionally, I would like to mention my friends Tak Yu, Withley, Hanky, Leroy, Stevie, Joe, Ben and Cheche for always helping me unwind and listening to me. Without any of these people I could not have achieved this.

Furthermore, I would like to thank Ronald Poppe, my supervisor, for the guidance, help and constructive criticism provided throughout the entire process. Additionally I would like to thank Fabian Mijsters and Ilse Smits for the help and support they provided to me during the data collection, and the theoretical discussions we had.

Contents

Chapter 1

Introduction

1.1 Problem Statement

There is clear evidence that pre-term birth is rising globally, yet the causes and implications are not fully understood. In 2010, an estimated 14.9 million babies were born prematurely, that accounted for 11.1% of the worldwide livebirths [1]. Given the rise of pre-term births, research for monitoring this sub-population of infants has become a widely research topic. Preterm infants are more likely to develop a myriad of developmental disorders compared to full-term infants [2], hence they require extensive monitoring. This monitoring is often performed by Neonatal Intensive Care Units (NICU), which provide a controlled environment for them to develop in. However, methods for monitoring infant health and development can be obtrusive as they rely on needles or electrodes, which are often painful or cause damage to their thin skin [3]. Therefore unobtrusive methods have become a welcomed and necessary tool for monitoring preterm infants in the NICU. Video-based methods are a relatively cheap and accessible type of unobtrusive monitoring methods that can be used to track and capture a wide variety of vital and behavioral signals. These methods are used in monitoring systems to detect seizures, hunger, sleep or discomfort and have an immeasurable value in medical applications. Movement and poses, which are what these methods seek to capture, are helpful indicators for tracking a wide range of developmental milestones of infants as well as their overall health. However, these methods require constant and active human participation which is generally not possible. Due to the large number of patients that are often in NICU, medical staff cannot monitor infant movements manually for large periods of time. Therefore recent work [4, 5] has aimed towards developing monitoring systems that are able to track and monitor the movement of infants in video through pose estimation. A wide range of state-of-the-art pose estimation methods are predominantly trained using only adults as their target domain, therefore when applied to infants these methods perform drastically worse [6]. The reason for this drop in performance is due to the fact that infants and adults have sig-

nificantly different body compositions [6, 7]; this is even more pronounced with neonates. Researchers such as Hesse et. al. [4,5] successfully modify and retrain state-of-the-art human pose estimation methods for the infant domain, however they deal with healthy infants in simple environments. For preterm infants, who are often placed in NICUs, the assumptions made by these methods are not met and pose estimation becomes a significantly more challenging task.

Heavy occlusions, both spatial and temporal, occur frequently due to movement and the presence of medical equipment, blankets, and other objects. This particular challenge is specially present for the preterm infant population and has not been fully addressed in the computer vision field. Additionally, lack of accessible infant data makes it difficult to collect and create diverse datasets for deep learning systems which are known to be data-reliant. Often, infant data is highly sensitive due to privacy concerns and even more so for the subpopulation of preterm infants as video recordings occur inside NICUs for medical purposes.These are critical challenges for developing computer vision systems to track infant movement in the medical domain and their solutions are not trivial. Therefore, this research proposes the extension of current state-of-the-art infant pose estimation from general population infants to preterm infants. Furthermore it aims to contribute to the infant pose estimation literature by directly addressing the challenges often present in this domain. In particular, it aims to create a network for preterm infant pose estimation to deal with occlusions in video.

Additionally, in order to specifically tackle the occlusions often present in the target domain, an occlusion augmentation technique for training data will be implemented. The intuition behind using occlusion augmentation is that by artificially occluding poses in video, we can regularize the pose estimator, and guide it towards not relying on a few joints and thereby robustly deal with heavy occlusions. The aim is to produce more temporally and spatially consistent results.

The ideal computer vision system for monitoring must be unobtrusive, cheap and easy to set up. Therefore the data that will be used to validate the proposed model will be video from a single RBG-D camera. The camera is set up at the side incubator of the infant in such a manner that it is does not interfere with the tasks of the caretakers.

1.2 Research Focus

This research assumes a supervised learning approach. It is focused around developing a deep learning approach to predict the poses of NICU infants from video footage. There are a several steps required in order to successfully create such an approach. Given our target domain, we must first develop a pipeline for generating synthetic data. One of the most crucial steps for pose estimation is the selection and use of annotated data. Annotated data of poses is required to

guide the loss function of the system towards learning a set of model parameters. For a standard human pose estimator, annotated data comes in 2D and/or 3D data and there a wide variety of paradigms to choose from. There are few annotated synthetic datasets for learning infant pose and shape parameters, such as MINI-RGBD [8], and SyRIP [7]. However preterm infant data is limited and the movement of healthy infants and pre-term infants is significantly different. Often 2D data annotation is time consuming task, especially when it is often heavily occluded. Synthetic data can often help in with this problem by providing 3D and 2D ground truth data, allowing us creating more abundant and diverse datasets to improve performance and robustness.

By following the methodology proposed by SMIL [5] and SMPLify [9], we can fit a volumetric model to preterm infant data and use it to generate synthetic 3D data. Given that the target domain generally often offers little to no data, synthetic data can successfully fill in the gaps missing in real data. The data used to augment pre-term infant movements is provided by the Utrecht University Medical Center. The reason for selecting a volumetric model such as SMIL over kinematic and planar models is due to its ability to capture both pose and shape information, which allows us to create highly realistic data.With the real and synthetic data generated, we can train a pose estimator using a hybrid dataset. The contributions of this research is two fold: (1) The application of the SMIL body model to the preterm infant domain with the capability of generating synthetic data of preterm infants, and (2) the development of an infant pose estimator that is capable of estimating pose of occluded pre-term infants.

1.3 Research Questions

The current state of the art has clearly indicated that we can model infant bodies movements appropriately with relatively small data using RGB-D data. However the same has not been attempted on preterm infant data. Preterm infants have slightly different bodies and the data is more incomplete due to the high amounts of occlusion present. Given that the aim is to develop an accurate body pose estimator for clinical application in NICUs the leading research question of this paper is the following:

(1) Can preterm infant poses be accurately estimated under occlusions in controlled NICU environments?

In order to verify whether the poses can accurately be estimated the MSE (Mean Squared Error), mAP (mean Average Precision), mAR (mean Average Recall) and PCK@0.2 (Percentage of Correct key-points) metrics will be used. However, to effectively answer (1), a series of sub-questions (2),(3), must first be addressed. SMIL [5] has demonstrated that infant bodies can be learned and

modeled. Additionally, they demonstrated that realistic synthetic data can be generated from such models in order to increase data diversity. In order to deal with the small data problem present in the preterm infant domain, this paper seeks to answer the following sub-question:

(2) Can a preterm infants movement be modeled in order to learn its pose parameters?

The intent behind such a question is to verify whether synthetic data of preterm infants can be generated. Given the current state of the field, it should be possible. In order to answer this research question, we will use the MSE error between the re-projected 3D joints and the ground truth 2D data to verify it, an MSE of less than 10 pixels should be acceptable.

The focus of this paper is primarily centered around estimating the poses of preterm infants and tackling the challenges that come with this domain, one of these challenges in particular are occlusions. Occlusions can be quite challenging to deal with, therefore it is necessary that techniques that aid in mitigating their impact on pose estimation results are studied. Therefore this paper aims to further study the impact of said techniques on performance and address the following sub-question:

(3) Can occlusion augmentation techniques used during training aid to minimize the errors for preterm infant pose estimation?

In order to answer question (3), this research will use the similar evaluation metrics as question (1): MSE and PCK. These metrics will be used to compare the performance of the network with and without the occlusion training. The research approach will seek to answer (2) and (3) in order, as they provide a solid foundation for the required steps that must be followed. Whether the preterm infants pose parameters can be modeled will validate the methodology for the synthetic data generation and answer sub-question (2). As for the occlusion augmentation sub-question (3), it will aid us in identifying whether these techniques are important for training deep learning models for the NICU domain. The main research question (1) will be answered once the sub-questions have been addressed and their respective techniques are applied to train an infant pose estimator.

Chapter 2

Literature Review

2.1 Preterm Infants

The term preterm infant refers to infants that are born before 37 completed weeks of gestation. There are categories of preterm birth based on their gestational age; extremely preterm (less than 28 weeks), very preterm (28 to 32 weeks) and moderate to late preterm (32 to 37 weeks). Depending on their gestational age, preterm babies have different physical characteristics that change over time. Table 2.1 provides an overview of the physical characteristics, relevant to the scope of this paper, of preterm children by their gestational age. Figure 2.1 shows an example of preterm infants of different gestational ages. As seen in Figure 2.1, the younger the infant the more red the skin is. The poses these infants are in provide clear examples for the information presented in Table 2.1; where the extremely preterm infant has an extended pose, the very preterm infant displays leg flexion, and the moderate preterm infant displays full limb flexion. The general trend seen in Figure 2.1 is that the physical appearance of preterm infants develops relatively fast.

Preterm infants are at a higher risk of developing serious disabilities, life threatening conditions and face a higher mortality rate due to their early birth; the earlier the birth the higher the likelihood [12]. Cerebral palsy, developmental delays, cognitive and physical disabilities are common conditions that preterm infants are likely to develop [13]. Additionally, preterm infants are more likely to develop sleep, attention and temperament dysfunctions than full-term infants [14]. These dysfunctions could have long-term consequences such as reduced psychomotor and reduced cognitive capabilities. This population is extremely sensitive; they require routine medical care and health monitoring. Currently, in high-income countries, neonatal intensive care units (NICU) provide a safe and controlled environment that allow us to obtain measurements that can be used to monitor preterm infants. Given that preterm infants are being taken care of in a medical environment, tubes, blankets, diapers and medical equipment are

Gestation	**Extremely Preterm**	**Very Preterm**	**Moderate/Late Preterm**
Skin	Very thin, gelatinous, dark red	Medium thin, pink	Thick skin with cracking, pale pink color all over ears, lips, palms and soles
Length	31 to 36.5 cm	36.5 to 42 cm	36.5 to 49 cm
Head circumference	21 to 26 cm	26 to 29.5cm	29.5 to 33.5 cm
Ears	Shapeless, little to no cartilage	Shaped, some cartilage is present on edge of ears	Shaped, Cartilage is mostly present
Eyes	Eyelids may be fused or partially open, absent or infrequent eye movements	Open eyes, increased eye movement	Open eyes, frequent eye movement
Posture	Extended, uncoordinated movement	Some flexion of legs, semi-coordinated movement	Flexion of limbs, coordinated movement
Musculature	Little to none subcutaneous fat, thin abdomen	some subcutaneous fat, thicker abdomen	pronounced abdomen, present subcutaneous fat

Table 2.1: Physical characteristics of preterm infants by gestational age. These physical characteristics are selected based on the scope of the paper. The characteristics are Skin, Eyes, Posture, Ears, Musculature, Length, and Head circumference. Data is compiled from Lissauer et al [10] and Fenton [11].

(a) Extremely Preterm (28 weeks)

(b) Very Preterm (32 weeks)

(c) Moderate preterm (36 weeks)

Figure 2.1: Preterm infants of different gestational ages.

likely occlude large sections of their body; as seen in Figure 2.1.

2.2 Preterm Infant State Classification

Preterm and term infant state monitoring is still an ongoing area of research. Non-contact and unobtrusive methods are of particular interest given that in preterm infants the epidermis is 2-3 layers thick and barely has a protective outer-layer [3]. Given the fragile skin, even attaching electrodes to the skin is considered too obtrusive; this greatly reduces the methods that can be applied. There are three types of measurements used for monitoring: polysomnography (PSG), polygraphy and behavioral measurements. These measurements can be used in three types of methodologies: PSG methods, behavioral methods or hybrid methods. PSG methods combine several polygraph measurements - such as heart rate and respiration - to determine the state of the patient in combination with PSG measurements such as EEG, ECG, and EOG signals. Meanwhile behavioral methods use physical cues such as motion and facial expressions; these are preferred for preterm infants due to their non-obtrusiveness.

This has sparked interest on improving and creating deep learning (DL) methods to collect both behavioral and PSG measurements that can be used for monitoring. In particular, video-based methods have recently become an interesting avenue for unobtrusive measurement collection and other clinical applications. However, the available DL models often under-perform in the infant domain due to the difference in body composition between infants and adults.

2.2.1 Video-based Methods

Video-based methods, using standard or depth cameras, are have become more accessible in recent years. They have seen a surge of uses in a wider variety of applications due to the possibility of extracting multiple signals from a single source and their non-obtrusiveness. Heart rate (HR) measurement has been

Figure 2.2: Example of an unobtrusive set-up for RR monitoring. Camera is highlighted with a red circle. Image was retrieved from Rossol et. al. [24].

a widely researched topic for video-based monitoring, producing a variety of methods [15–22]. Verkruysse et. al. [16] were the first to be able to remotely measure HR and Respiration Rate (RR) using a digital camera by applying digital filtering and spectral analysis using ambient light. Aarts et. al. [17] tracked a region of interest and further applied this technique under similar lighting conditions and successfully measured HR of newborn infants in the NICU from the color changes in the skin. Figure 2.2 demonstrates an example of a video-based set-up in a NICU. Koolen et. al. [18] and Naji et. al. [23] were able to detect RR from video data of neonates during deep sleep by extracting optical flow information from Eulerian magnified videos.

Heinrich et. al. [25] proposed to analyze body movement by employing a spatiotemporal-based recursive search on video motion as a replacement for current motion-based sleep (actigraphy) monitoring methods. Long et. al. [26] further developed this concept by classifying video-based actigraphy signals using a Bayesian-based linear discriminant classification model for monitoring wake and sleep in healthy infants. Cabon et. al. [27] used a multi-modal approach to estimate the sleep state of newborns in the NICU by creating feature descriptors from audio, motion analysis, and eye state estimation. For both audio and movement, they were interested in globally capturing the amount of activity at a certain time and creating a continuous signal. To capture movement activity, they compared subsequent frames and calculated the difference between pixels.

More recent research [4, 5, 28] has focused on leveraging infant pose estimation in order to capture, track, and monitor movement of infants in video. These methods provide a solid foundation and proof that collection and monitoring of PSG signals from video is valuable and practical in a clinical setting, particu-

larly in the NICU and developing approaches to tackle the challenges present in this domain are necessary. Therefore we aim to further iterate on current methodology for preterm infant pose estimation.

2.3 Human Pose Estimation

Human Pose Estimation (HPE) is a technique that allows us to measure a wide variety of behavioral signals from videos, with the goal of predicting the locations of body joints. HPE provides us with the tools to study a wide range of properties about the human and what it is capable of. It allows us to study the movement of humans, as well as to keep track of certain regions of interest in the body. Robust applications of HPE in the clinical domain, specifically for the preterm infant population, allows us to non-obtrusively monitor patients and to collect behavioral signals such as phasic twitching and movement. These signals can then be used to identify physiological states of patients such as hunger or pain, and for applications such as seizure detectors.

2.3.1 A taxonomy of Human Pose Estimation

HPE can be divided into two directions, 2D and 3D HPE. 2D methods try to predict the 2D spatial location of joints of human body key-points in images/videos. Similarly, 3D methods try to predict these spatial location of joints in 3D space. Unlike 2D methods, 3D methods can provide extensive 3D structure information about the body and movement [29]. DL methods have shown great promise for HPE in both 2D and 3D; DeepPose [30] brought forward a paradigm shift. Classical methods which rely on hand-crafted feature extraction and sophisticated body models to obtain local representations and global pose structures [29,31] have been outperformed by DL frameworks and in recent years, DL has been the backbone of progress in the field. These methods are aimed towards either single-person or multi-person pose estimation. Given that the target domain for this research is preterm infants in the NICU, the research of this paper is centered around single-person 2D HPE. Methodologies in single-person 3D HPE can be grouped depending on whether they use a human body model or not [32]. Additionally these DL HPE frameworks can be further divided into an array of categories: bottom-up vs. top-down, regression-based vs. detection-based and one-stage vs. multi-stage [29, 31]. Each of these frameworks have their advantages and limitations.

Model-free vs. Model-based

Model-free methods generally learn a mapping between images and 3D human poses by directly estimating joint locations (direct regression), or by exploiting

2D pose estimation and use the intermediately estimated poses to predict 3D poses (2D-to-3D lifting). Recent work by Pavlakos et. al. [33] proposed a multi-stage framework to improve the performance of direct regression methods. They propose a natural volumetric representation by performing a discretization of the 3D space into voxels. Additionally they propose a gradual refinement scheme called coarse-to-fine prediction scheme. It was identified that naively stacking multiple components gave diminishing returns, therefore they decided use different resolution targets at different training stages. In the first iterations they supervised the training with lower resolution targets; they gradually increased the resolution of targets in later iterations. By doing so, they were able to stack components and avoid the overfitting and dimensionality issues that are associated with such an approach. Sun et. al. [34] proposed a structure-aware approach, where they used a bone-base representation instead of a joint-based representation. This allowed for the encoding of long range interactions between bones, this resulted in increased stability within poses.

2D-to-3D lifting approaches tend to outperform direct-regression methods, and are the more commonly used of the model-free methods. They benefit from the progress made in 2D pose estimation [29] as they use state-of-the-art 2D HPE networks. The 2D-to-3D lifting step has been predominantly done by training neural networks. Martinez et. al. [35] were one of the first to train a simple residual CNN to regress 3D joint locations from 2D and predict depth. They provided a baseline for simple 3D HPE and achieved state-of-the-art performance at the time. More recent work, such as Pavllo et. al. [36] developed Temporal Convolutional Networks (TCN), which used a encoder-decoder architecture to perform a semi-supervised training scheme by using a 2D-to-3D pose estimator as an encoder and a projection layer as a decoder. Zhou et. al. [37] developed a weakly supervised transfer learning approach that used mixed 2D and 3D labeled data to train a 2D pose estimator and a 3D depth regression sub-network both sequentially and separately. By presenting 2D heat-maps alongside with intermediate feature representations from the 2D pose estimator to the depth regressor they were able to extract semantic information that served as additional cues for 3D pose recovery.

Model-based methods, on the other hand, employ parametric body models to estimate the shape and pose from images [31]. The human body models used in are either kinematic, planar, or volumetric. Most work has centered around kinematic and volumetric models. Kinematic models represent the body as a structured graph, allowing for a very simple yet flexible representation. However, due to their simple nature they provide no information regarding shape. Zhou et. al. [37] embedded a kinematic object model in a ResNet for general object pose estimation. By doing so, they were able to constrain both the orientation and rotational properties of the model and produce more accurate estimations. Mehta et. al. [38] fitted a kinematic skeleton model against 2D and 3D pose predictions from a single RGB-D image. They were able to constrain the output of their pose estimators and produce temporally stable joint angles

of a metric global 3D skeleton in real time.

Volumetric models, unlike kinematic models, are capable of recovering high-quality meshes and thereby provide additional shape information of the body. One of the most commonly used volumetric models is SMPL. Developed by Loper et.al. [39], SMPL is a skinned vertex-based statistical model that is able to represent a wide variety of body shapes in natural poses. In the original paper, they are able to train the model from a large variety of pose-aligned 3D meshes of different people in different poses. Bogo et. al. [9] developed SMPLify, where they estimated the 3D pose from a single unconstrained image. They used DeepCut [40] in order to predict the 2D joint locations and fitted the SMPL model to generate the 3D pose and minimize the re-projection error. In more recent work, Varol et. al. [41] used SMPL in order to generate 3D body meshes for a synthetic human database. Similarly to previous work, they trained an Hourglass [42] backbone for 2D pose prediction and body part segmentation, and depth estimation. They leveraged the 2D and 3D data generated with SMPL to train the network and demonstrated that synthetic data can be effectively used to learn 2D/3D poses. Other volumetric models have been developed, such as the Cylinder Man Model by Cheng et. al. [43], which was used to create 3D ground truth data for occlusion data augmentation.

Top-down vs. Bottom up

Top-down frameworks generally have two steps, the human detection step and the pose estimation step. Top-down frameworks use a person detector to identify where in the image humans are present and a single-person pose estimator is ran for each individual found [29]. Bottom-up frameworks do not require a person detector, but rather predict all joint locations in an image and then treat it as a clustering problem. As Ning et. al [44] indicate, an advantage of top-down approaches is the fact that they disassemble the task into multiple comparatively easier tasks. Having an object detector that is trained for detecting hard candidates improves the performance of the pose estimator by focusing the regression space. However, this can inversely be seen as a negative quality given that the results of top-down frameworks are heavily dependent on the quality of the person detector results [45]. In complex environments where occlusions are common phenomena, top-down approaches can suffer from early-commitment issues due to the person detector and fail to recover [46]. Figure 2.3 provides a visualization of how bottom-up and top-down frameworks generally function.

Regression vs Detection

DL regression-based frameworks were pioneered by DeepPose [30]. Regression-based methods attempt to solve a highly non-linear problem by directly mapping the input image to the coordinates of body key-point joints or to the parameters of human body models [31]. These models are primarily used in 3D.

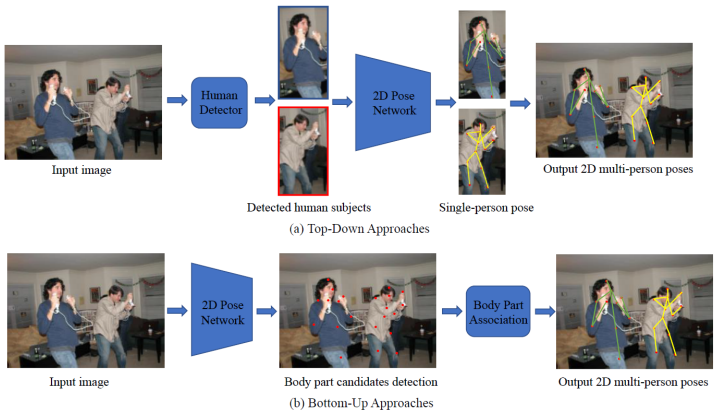


Figure 2.3: Visualization of generic top-down and bottom-up methods. Retrieved from Zheng et. al. [29].

Regression-based methods are fully differentiable, however they lack inherent spatial generalization [47]. Unlike regression-based methods, detection-based frameworks try to predict the approximate spatial locations of body key-point joints [29], these were pioneered by Tompson et al. [48]. Detection methods do this by generating a likelihood heat-map for each joint and using the point with the maximum likelihood as location of the joint [49]. In practice, detection methods have shown to outperform regression methods and are more common in 2D single-person pose estimation; the dense pixel information appears to facilitate the heat-map supervised learning. In 3D HPE however, regression-based methods are more common given that heat-map representations are more computationally expensive in 3D.

Detection-based methods are not free of limitations. Output heat-maps are of lower resolution due to the down-sampling performed in neural networks which introduces quantization errors [49]. Furthermore, these methods tend to perform worse with low resolution inputs [50]. In order to address this problem, Cheng et. al. [51] used a feature pyramid consisting of feature map outputs from their previous work [52]. HRNet [52] is used as the backbone for HigherHRNet. This backbone produces high-resolution representations of the input, obtained by applying parallel and repeated multi-resolution feature fusions. Rather than trivially applying a Gaussian filter to smooth lower resolution features, they up-sampled the highest-resolution outputs of HRNet by passing them through a transposed convolution. By doing so they were capable of achieve state-of-the-art results on low resolution images. Detection-based methods tend to be sensitive to body occlusions and background complexity [31] and can have high-memory requirements due to the nature of heat-maps, making them difficult to implement in systems with low computational and memory resources [45]. Regression-based methods have the advantage of generally being end-to-end

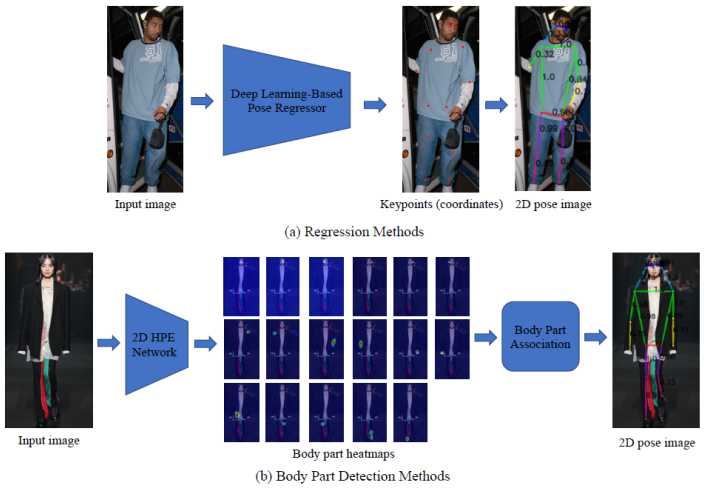


Figure 2.4: Visualization of generic regression and detection-based methods. Retrieved from Zheng et. al. [29].

and producing a continuous output, which is not possible with detection-based methods due to the non-differentiable nature of joint coordinates in heat-map representations. Figure 2.4 visualizes how regression and detection-based methods differ in terms of what is learned.

In recent years, some methods [47, 49, 50] have attempted to bridge this gap between regression and detection methods in order to exploit heat-map representation and regression advantages and produce end-to-end differentiable frameworks, resulting in detection-based frameworks that are as flexible as their regression-based counterparts. Diogo et. al. [50] and [47] proposed to use the soft-argmax function for HPE, which could be implemented as a CNN layer in order to convert 2D feature maps directly into 2D joint coordinates, thereby bridging the regression and detection based methods. The soft-argmax function makes detection-based methods fully differentiable by performing a weighted average of the confidence heat-maps. By doing so, confidence heat-maps can be converted to spatial coordinates. They were able to train end-to-end networks to learn heat-map representations indirectly, achieve comparable results with state-of-the-art detection-based approaches and directly addressed the poor performance of detection-based methods in lower-resolution settings. Sun et. al. [49] further iterated on this technique by applying it to both 2D and 3D training data and studied the effect of resolution and network capacity with integral regression, and its impact on multi-stage framework performance. They found that multi-stage networks with integral regression improved performance as stage increases. Furthermore, integral regression significantly alleviated quantization errors and are significantly less affected by resolution compared to heat-map

based methods. Lastly it was found that in terms of parameters and complexity, smaller networks with integral regression outperformed larger networks without it.

One-Stage vs. Multi-Stage

One-stage frameworks aim to map input images to human poses by employing end-to-end networks. Generally one-stage frameworks tend to be regression-based and top-down [53]. Multi-stage frameworks predict poses in multiple stages and generally have intermediate supervision [31], which is done in order to address the vanishing gradient problem [46]. Multi-stage frameworks can take many forms, however the foundation remains the same - stack multiple networks on top of each other. Newell et. al. [42] proposed an encoder-decoder architecture coined Hourglass, since then this architecture has been a common choice for a variety of Multi-stage frameworks [33, 43, 54–56]. Hourglass-based networks apply intermediate supervision and repeat bottom-up and top-down processing to capture information at every scale. One-stage frameworks generally tend to be easier to train, however they have less intermediate constraints. Multi-stage networks are seemingly more suited to the task as they provide more flexibility, and benefit heavily from intermediate supervision [36].

In the following sections an overview of the challenges HPE faces will be presented, with a focus on the challenges imposed by the clinical and infant domains, followed by a summary of current infant pose estimation literature.

2.3.2 Human Pose Estimation Challenges

The challenges present for HPE primarily arise from the human body and the complex environments in real-world settings. The body is challenging to learn due to the fact that it is flexible, joints can be interdependent and have high degrees of freedom, which allow for complex and rare body positions [31] - this is particularly true for the bodies of infants [5]. Additionally, the body is capable of producing non-linear motions and has high pose and appearance variance [32]. 3D methods suffer from the problem of depth ambiguity. This problem is ill-posed as multiple 2D poses from monocular videos can produce a single 3D pose due to depth ambiguities. Furthermore, this is an inverse problem given that during projection from 3D to 2D (real world to camera) a dimension is lost [29].

Occlusions

Dealing with occlusions (self-occlusions and foreground occlusions) and ambiguities are common challenges in HPE. It is common for bodies to cause self-occlusions, particularly in unrestricted environments. In the clinical setting heavy occlusions are a frequent challenge due medical equipment, medical staff,

the body, blankets and other objects [5]. Intensive care units tend to have heavily occluded patients due to the critical nature of their condition, the same holds for preterm infants in NICU. Therefore it is critical to develop systems that are capable of handling occlusions, caused by foreground objects or the body of the patient itself. Attempts have been made to directly tackle occlusions that are a common phenomena in the clinical domain. Achilles et. al. [57] developed Patient MoCap, a dataset of motion-captured 3D video data with a simulated blanket occlusions of varying degrees and trained a Recurrent Neural Network (RNN) on this dataset to directly regress 3D joint locations for human poses. They used an RNN, which at the time was a state-of-the-art DL method to capture spatio-temporal information. Given that occlusions do not occur in a single frame, spatio-temporal information is incredibly useful as it allows us to smooth and obtain more consistent pose predictions. Older work generally used RNNs to encode temporal information, however more recent work has moved on towards using Temporal Convolutional Networks (TCNs) [36].

Temporal Convolutional Networks (TCN) were developed by Pavllo et. al. [36]. TCNs [36] are multi-stage, 2D-to-3D lifting networks which use a 2D pose predictor to generate a sequence of 2D joint key-point poses from video. For each frame, the joint coordinates are concatenated and a temporal convolution is applied. They apply 1D convolutions over a series of ResNet-style blocks to increase the receptive field exponentially and capture temporal dependencies. The convolutional nature of these networks allows parallelization over both time and batch, offering precise control over the temporal receptive field and mitigates vanishing and exploding gradient problems.

Cheng et. al. [43] developed an occlusion-aware network that directly addressed self-occlusions by deploying a multi-stage framework. In addition they develop the Cylinder Man model, a volumetric model for occlusion reasoning that is used as a heuristic to map 3D ground truth points to 2D heat-maps. First a detection-based 2D pose estimator and an optical-flow consistency constraint are deployed in order to obtain joint confidence maps. By filtering out non-reliable estimations of occluded key-points they produce incomplete but correct 2D pose key-points that are fed to two TCNs, one in 2D and another in 3D. The reasoning behind this is that by giving an incomplete but reliable set of points, occlusions can be explicitly modeled. The 2D network takes as input the 2D pose key-points, while the 3D network utilizes a pair of 2D pose key-points and 3D ground truth key-points fitted to the Cylinder Man model. A limitation of this approach is that the cylinder man model can only model self-occlusions and does not explicitly address foreground occlusions. Furthermore the method does not perform well with long-term heavy occlusion. This work was further iterated on by Wang et. al. [58], they used the method described by [43] and sought to train the 2D TCN to recognize the ground truth occlusions by explicitly adding them to the loss function and created the Boxed Man model, an alternative to the Cylinder Man model that is less computationally expensive and has a higher tendency to mark joints as occluded. The higher tendency for

occlusion labeling of the Boxed Man model works as a dropout or regularization technique and helps guide the network such that it does not overly rely on a subset of joints. The results achieved were comparable, albeit slightly better than the original work.

Other recent work has focused primarily on dealing with foreground occlusions. Foreground occlusions occur when other bodies, viewing angles, or objects prevent us from (fully or partially) seeing parts of the target body. In order to tackle occlusions in images Zhou et. al. [59] developed an occlusion-aware siamese network. They leverage an attention mechanism to remove the interference caused by occlusions. By artificially occluding images, they train a sub-module of the network to predict heat-map joint locations as well as occlusion maps and erase the contaminated features in a multi-task fashion. A reconstruction sub-module is trained to reconstruct the images by providing the feature erased image with the original non-occluded image. However, occlusions usually do not occur in a single frame, rather they occur persistently across multiple frames [43]. In more recent work, Cheng et. al. [60] further focused on occlusion handling. They apply high resolution networks [61], which perform repeated multi-scale fusions to generate joint heat-maps, and concatenate them to obtain multi-scale features. These features are then given to an embedding network to generate low dimensional representations. By applying TCNs with multiple strides to the multi-scale features they are able to incorporate more spatial and temporal information to the predictions. This allows them to deal with pose estimation in multiple scales for both the temporal and spatial domain. Furthermore, they introduce a novel spatio-temporal pose discriminator to reduce the risk of generating unreasonable 3D pose sequences. After training the discriminator, they use it to produce a regularization loss that is used for the pose prediction training. Lastly they perform occlusion data augmentation, three types of occlusions are applied during training. The first type they use is frame-wise occlusions, where they randomly mask several frames in a video. The second type is point-wise occlusions, where they randomly set certain keypoint heat-maps to zero. The aim behind such an technique is to simulate that certain points are occluded. Lastly, they applied area occlusion by setting a virtual area to be occluded; every heat-map activation in this area is set to zero. This third augmentation aims to simulate occlusions that occur across bigger sections of the body.

Data Scarcity

DL has introduced its own set of requirements for HPE. DL methods require large amounts of data in order to learn the target domain [29], which is not readily available in certain domains. This is somewhat addressed by using pre-trained network layers as the backbone for feature extraction followed by fine tuning to the target domain [62]. The data requirement can be further alleviated by the introduction of data augmentation techniques. Although performing data

augmentation is not as good as having more real data, it allows us to increase the available data provided to the data-hungry networks. Data augmentation is particularly useful when little data is available. Occlusion augmentation is one of such methods, papers such as [43,57,59,60,63] use this technique. Others such as [5, 7, 9, 41, 42] use volumetric models to generate synthetic data. Generative Adversarial Networks (GANs), have also become a powerful tool for generating synthetic data [64–67].

Data used to supervise training requires annotation. For 3D HPE collecting accurate pose annotations is time-consuming and manual labeling is impractical, although this can be somewhat avoided with data augmentation techniques. As Zheng et. al. [29] indicate, creating 2D human pose datasets with accurate 2D annotations is more practical. Furthermore, 2D data sets are more inexpensive in terms of the equipment required to collect the data and the time spent annotating the data. Others such as [36] have moved to semi-supervised methods to further address this. By leveraging a high-performing 2D pose estimator, they predict 2D poses on unlabeled data and feed them to their TCNs to predict 3D poses. The 3D poses are then re-projected to 2D, the training then penalizes re-projected points that are far from the original predictions. Leveraging 2D image datasets for 3D HPE has become a wide-spread solution [68], hence the popularity of 2D-to-3D lifting methods.

For the infant domain, data is generally not available, as it is difficult or impossible to collect due to ethical concerns. As Hesse et. al. [5] point out, infant data is often of low quality, noisy and subjects have large parts of the body occluded. Additionally, many HPE methods do not work out-of-the-box with infants, as these tend to be only trained on adults. The body proportions of adults and infants and the versatility of their poses are significantly different [6, 7]. In the next subsection, we will discuss methods developed to predict the poses of infants and address the challenges in the domain.

2.3.3 Pose Estimation for Infants

Pose estimation for infants is relatively sparse, however it has the potential to aid in child monitoring in clinical environments and trials. In order to study the efficacy of human pose estimation methods on infants, Sciortino et. al. [6] developed an infant dataset that contained images with varying degrees of occlusion and truncation and studied the performance for state-of-the-art pose detectors. They were able to confirm that every detector performed significantly worse in the infant domain. Hesse et. al. [69] used 3D pose estimation in order to analyze the motion of infants with the intent to capture early signs of movement disorders. By using an RGB-D camera they are able to train random ferns to assign a body part to each input depth pixel. Further reiterating this idea, Hesse et. al. [70] achieved a better performance by applying a feature selection step prior to training. They remove redundant features by randomly generating a large set of features and evaluating their information gain on the training data

Figure 2.5: Example of a modeled infant using SMIL. Retrieved from Hesse et. al. [5]

once. Only features whose information gain is above a user-specified threshold are kept. Additionally they incorporated a kinematic chain reweighing scheme to constrain poses and identify misclassified pixels. After the development of SMPL, in a series of papers with a similar aim as [9], Hesse et. al. [4, 5] learned a 3D Skinned Multi-Infant Linear body Model (SMIL) from noisy, low-quality and incomplete RGB-D data of infants in a clinical environment. An example of the modeled data can be seen in Figure 2.5. They use SMIL to capture body pose and personalized shape, as well as face and hand landmarks. In order to do so they replace the SMPL mean shape with an infant mesh and scale the pose blend-shapes to infant size. Lastly they adjust the pose priors manually in order to prevent the model from explaining shape deformations with pose parameters. Additionally, they perform a case-study on general movement assessment of infants and demonstrate the capacity of SMIL to faithfully represent the shape and pose of infants. In more recent work, Huang et. al. [7] applied the SMIL model to generate a hybrid dataset of synthetic and real images of infants called SyRIP. Additionally they propose FIDIP, a fine-tuned domain-adapted infant pose estimation framework that is easy to integrate with any encoder-decoder pose model for training with hybrid datasets. This framework proposes the introduction of a domain confusion network during training, which shares the feature extractor from the pose estimation component of the encoder-decoder model and adds a domain classifier head. The aim of this classifier is to identify whether the image is synthetic or real. By doing so, the hybrid and real image domain are mapped in the same feature space after extraction.

Moccia et. al. [28] opted for estimating the 2D limb-poses of preterm infants in the NICU from depth images. Rather than estimating full poses, they focus on capturing information about limb movement as they are incredibly helpful predictors to diagnose cerebral illnesses. Inspired by [46] they build a bottom-up, multi-stage, detection-based framework where they use two consecutive CNNs. Their framework has one CNN for detecting joints probability maps and joint connections using part affinity fields, and another CNN for regressing joint positions. The detection network uses a classic encoder-decoder bi-branch

architecture, where the network receives as input 20 ground-truth binary detection maps from a video frame. The final output of the network is a confidence map of joint and joint connections. Similarly, the regression network receives as input the depth image and the output of the detection network, as output it produces joint confidence maps. They link joints with their joint connections by applying non-maximum suppression and exploiting the joint connection regression maps using a bipart matching approach. Moccia et. al. [71] continued their work by adopting temporal clips to encode temporal information to constraint poses and obtain more consistent results. They converted the framework from their previous work [28], and used 3D CNNs to encode temporal information.

Chapter 3

Synthetic Data Generation

The creation of synthetic data allows us to increase the amount of data that can be used to train and evaluate the proposed preterm infant pose estimation network. Infant Pose Estimation is a challenging task due to the nature of the domain. Infant pose data is often scarce and difficult to collect and distribute due to a myriad of legal and practical reasons. The limitations imposed by the domain create the challenges of data scarcity, lack of pose variability and reduced quality.

To address this, Hesse et. al. [5] developed SMIL, an adaptation of the SMPL [39] model trained with incomplete 3D scans of freely moving infants placed on their backs. The SMIL model is highly descriptive for body shape and pose and can be used to create synthetic data of infants. Huang et. al. [7] demonstrated that by using realistic synthetic data generated with SMIL we can address the challenges present in the infant domain and train better models. They were able to demonstrate that models trained on augmented hybrid data achieved better results than those trained only on real training data with limited pose quantity and variability.

Initially, it was expected that the annotated data collected for this research would contain the depth information of the scene. Lighting conditions, incubators, and practicality limitations in the NICU environment significantly decreased the quality and viability of using such depth recordings, therefore only 2D information was available from annotated data. This impacted the synthetic data generation methodology as it was no longer possible to follow the approach proposed by Hesse et. al [5] to learn pose and shape parameters of preterm infants from collected point clouds. In order to circumvent this limitation, the research opted to follow and alter the approach proposed by Huang et. al. [7] instead, which leverages SMPLify [9] to estimate the SMIL body parameters from a single 2D image. In the following sub-sections we will quickly go over SMIL and SMPLify in order to contextualize the modifications required for the preterm infant domain.

3.1 SMIL

SMIL [5] is a skinned vertex-based statistical model that is capable of capturing both infant shape and pose. This model contains three crucial components: a template mesh, a pose prior and a shape prior. The SMIL model was learned from 200K frames of 37 sequences of freely moving infants. The SMIL template mesh was created by using the same topology as SMPL, this template mesh contains $N = 6890$ vertices and $K = 23$ joints. Unlike SMPL, SMIL only has one mesh template for gender. SMIL only has a neutral gender due to the fact that gender does not have an effect on the body shape of infants. The model can be parameterized by two sets of coefficients, the pose coefficients $\theta \in \mathbb{R}^{3(K+1)}$ and the shape coefficients $\beta \in \mathbb{R}^{20}$. The pose coefficients describe the pose of the infant using an axis-angle representation, where each joint has 3 degrees of freedom (DoF) and each can be described by an axis angle that represents the relative rotation of body parts. Given that there are $3(K+1)$ pose parameters, 3 of these parameters describe the global translation γ of the pelvis (root of the kinematic tree), for a total of 72 pose parameters. The shape coefficients represent the proportions of the individual's height, head-to-body-ratio, as well as torso and limb length, fatness and thinness. The pose and shape prior are learned from registering the real world data to the template mesh. To learn the pose prior, the 200K poses were filtered by using a minimum difference threshold, resulting in a total of 47K different poses used. To learn the shape prior, a personalized shape unrestricted from the shape space of the model was created for each infant. These shapes are created by uniformly sampling 1 million points from fussion clouds, which are the union of 1000 randomly sampled point clouds with virtual points, and performing gradient based optimization to minimize the difference between the template mesh and scans. The shape prior is created by performing a weighted principal component analysis (PCA) on the personalized shapes with an iterative expectation-maximization approach and retaining the first 20 components. The template mesh is deformed to be the average of the personalized shapes. Therefore the result of the fitting the mesh of the SMIL model can be represented as $M(\beta, \theta, \gamma)$, where M represents the template mesh that is deformed given a set of θ, β, γ coefficients to fit the body of an individual in a video.

3.2 SMPLify

SMPLify [9] is a 2D-to-3D lifting approach used to estimate the θ, β and γ parameters of models that share the topology of SMPL, such as SMIL. SMPLify assumes that the person in the frame is parallel to the image frame, meaning it uses a weak-perspective camera model C which has a set of parameters K. By calibrating the camera we are capable of obtaining all the required intrinsic parameters of the camera model, however due to our non-static monocular set-up of the data collection and the fact the camera is not calibrated before each recording, the extrinsic camera parameters are unknown. This adheres

to the assumptions made by SMPLify that both camera translation and body orientation are unknown and the focal length is known, making it the optimal approach for fitting the SMIL model to our data.

The first step is to estimate the camera parameters. SMPLify only estimates the camera translation, camera rotation is ignored given that it assumes the person in the frame is parallel to the image frame. The camera translation is initialized to be the γ of the SMIL model. In order to estimate the depth (or z-coordinate) of the camera, SMPLify calculates the ratio of similar triangles. SMPLify does so by dividing the mean euclidean distance of the 3D joint locations d_{3D} by the mean euclidean distance of the 2D ground truth landmark annotations d_{2D} of the shoulder and hip joints. The ratio of similar triangles is then multiplied by the focal length K_f. The estimated depth parameter is then as follows:

$$K_{t_z} = K_f \frac{d_{3D}}{d_{2D}} \tag{3.1}$$

SMPLify then refines the camera translation K_t by minimizing the objective function (2) from Bogo et. al. [9]. This equation is the weighted robust distance between the 2D landmarks and the corresponding 2D projections the SMPL model joints. For a more precise description refer to their work; the only joints that are relevant for the purpose of fitting the camera are the shoulder and hip joints. With the camera parameters estimated, SMPLify fits the model in a staged approach by minimizing Eq. (3.2) derived from Bogo et. al. [9]. The E_{data} term penalizes the weighted 2D distance between the joint landmarks and the corresponding projected SMIL joints. The pose prior term, E_θ, penalizes improbable poses given the θ coefficients, and similarly the shape prior E_β punishes improbable body shapes given the β coefficients. It is important to note that the pose and shape priors of SMPLify are substituted by those proposed by SMPLify-X [72] for the E_θ and E_β terms presented in Section 3.3.1. Lastly, Bogo et. al. [9] introduce the angle term E_a to heavily penalize unnatural knee and elbow rotations; where unnatural rotations are positive rotations. E_a is the sum of the exponential of the rotations of knees and elbow joints. Lastly, E_{sp} is the interpenetration term that uses capsule representations of the body to punish shape intersections; rather than using SMPLify's version of this term, we opted to use the improved interpenetration term proposed by Pavlakos et. al. [72].

$$E(\theta, \beta) = \lambda_{data} E_{data} + \lambda_\theta E_\theta + \lambda_\beta E_\beta + \lambda_a E_a + \lambda_{sp} E_{sp} \tag{3.2}$$

3.3 Data Generation Procedure

As previously indicated, the body and pose parameters for SMIL were estimated by deploying a modified SMPLify fitting procedure which uses a gradient-based optimization. As recommended by [72], we use a L-BFGS with strong Wolfe

line search optimizer, with a learning rate of 0.1, allow at most 30 optimization iterations with a gradient and loss tolerance rate of $1e-9$. In the following sections, the energies that were minimized for the fitting procedure will be explained, as well as the modeling choices behind them.

3.3.1 Fitting

The SMIL model is initialized by setting the θ coefficients to the mean of the SMIL pose prior. In order to register the SMIL model to the sequence, the best initial frame is selected. Similarly to Hesse et. al. [5], an automatic method for detecting the best frame is applied where the best frame is defined to be the one that contains the most visible body part segments. We can define a body part segment b to be the set of two annotated landmarks, and s is set of all body part segments in the landmark annotation of a frame. Given this notation, a sequence of landmark annotations S can then be defined simply as the set of all landmark annotations s of an individual video. Each landmark has a 'confidence' score c, the confidence score depends on the visibility label from the annotated data. The mapping from label to confidence score is as follows: not visible/occluded = 0 and visible = 1. The visibility of a segment is calculated to be the euclidean distance between the two points of a segment multiplied by their confidence value. The equation to find the best frame F, from a sequence of landmark annotations S can be defined as:

$$F = \arg\max_{s \in S}(\sum_{b \in s} d(l_1, l_2) * c_{l_1} * c_{l_2}), \text{where } l_1, l_2 \in b. \tag{3.3}$$

The best frame is used to predict the camera parameters for SMPLify. The camera parameters are then kept fixed throughout the entire registration procedure for each sequence. In order to register the model to the sequence and fit SMIL to the infant, we minimize an objective function similar to the one function proposed by Bogo et. al. [9] for each frame, we optimize the following energy w.r.t the pose θ and shape β parameters:

$$E(\theta, \beta) = \lambda_{data}E_{data} + \lambda_\theta E_\theta + \lambda_\beta E_\beta + \lambda_a E_a + \lambda_{table}E_{table} + \lambda_{to}E_{to} + \lambda_h E_h + \lambda_{sp}E_{sp} \tag{3.4}$$

Eq. (3.4) is a domain adapted version of SMPLify's objective function (3.2). The additional terms $\lambda_{table}E_{table}$, $\lambda_{to}E_{to}$ and $\lambda_h E_h$ are introduced to further constraint the optimization energy and enforce penalties for movements and poses that are deemed impossible for the preterm infant domain. In the following paragraphs the intuition behind the changes and additions introduced to the objective function will be expanded upon.

Pose Prior Term

Originally, the pose prior of SMIL was used to penalize unlikely poses due to the fact the prior was learned from a similar domain. A problem with using

such a prior is the fact that the infants from which the SMIL prior is learned are much older. By performing a qualitative evaluation on the results early on, it became apparent that the SMIL prior was not representative of the preterm infant domain. The infants used to learn the SMIL prior are capable of striking a very different distribution of complex poses, in particular with the limbs than those a preterm infant can perform. Therefore we opted to use the pose prior of adults from SMPLify. The methodology originally proposed by Huang et. al. [7] did this as well, however in that case it appears to be a design oversight rather than a choice. The pose prior term used is a Gaussian Mixture from a MoCAP dataset and is explained by equation (5) in Bogo et. al. [9].

Shape Prior Term

With the improvements of Pavlakos et. al. [72] and following the methodology of Huang et. al. [7] the shape prior term is set to be the squared Mahalanobis distance between the β coefficients and the shape prior of SMIL. Given that the shape prior is constructed to have a mean of 0, the shape term is simply as follows:

$$E_\beta = ||\beta||^2 \tag{3.5}$$

Angle Prior Term

Given the axis-angle representation of the joints, the formulation presented by Bogo et. al. [9] allows us to heavily punish implausible joint rotations. Since the axis angle has a range of $(-2\pi, 2\pi)$, we can punish joint rotations in specific axes and directions. For example, the elbow should never turn outwards to explain a wrist rotation as that would not be physically plausible given the elbow joint. This term guides the optimization towards modifying other pose and shape parameters to explain the location of the ankle and wrists in the image. The equation for the angle prior term proposed by Bogo et. al. [9] is the following:

$$E_a(\theta) = \sum_{i \in (elbows, knees)} e^{\theta_i} \tag{3.6}$$

Table Term

Given the non-static monocular set-up of the data collection, a table term used to enforce that the infants are laying on an incubator, like the one proposed by Hesse et. al. [5] was not possible. However, given the preterm infant domain certain constraints could be placed on the body joints. The infants from our collected data are always on their backs, laying inside of an incubator. Inspired by the angle prior term, the E_{table} term punishes shoulder and hip rotations around the x-axis that would place the arm or legs below the torso. Although this does not fully remove some of the fitting artefacts, it helps to guide the

optimization and simulates a virtual table. This term is identical to the one described in Eq. (3.6) with the exception that $i \in (shoulders, hips)$.

Torso Term

In order to further constraint the fitting procedure, we enforce the limited strength and mobility of preterm infants to punish implausible poses. The E_{to} term punishes large differences between the rotations of the hip and neck joints around the x-axis, meaning the torso should remain relatively straight, the infant should never have its neck higher or lower than the hip. The intuition behind this term, is that preterm infants are not able to move as freely as full-term infants and hence are not capable of raising their torso or hips in the same manner. This term further aids to enforce that the infant should be laying flat on its back. The term can be formalized as the following:

$$E_{to}(\theta) = e^{||\theta_n - \theta_h||} \tag{3.7}$$

where θ_n and θ_h refer to the neck and hip joint pose parameters we want to constraint.

Head Term

The head term is introduced in order to limit two particular rotations the infant cannot perform. The first rotation we want to punish is the movement of the head around Z-axis regardless of direction; the second rotation we punish is the neck rotation around the X-axis. Once again these movements are deemed to be highly unlikely for infants to perform and help to enforce the restricted poses of preterm infants during the fitting procedure. Given that we want to constraint these two movements, the E_h is the following:

$$E_h = E_l + E_u \tag{3.8}$$

Where E_l punishes the movement around the Z-axis and is formalized by equation (3.9). By using the absolute value of the θ_n for the neck joint Z-rotation we can produce a similar behavior to that of a quadratic and punish movements regardless of direction. The term E_u punishes the movement on the X-axis and is formalized by Eq. (3.6).

$$E_l(\theta) = e^{||\theta_n||} \tag{3.9}$$

3.3.2 Term Weights

As previously mentioned, the registration is executed in a staged approach, following the advice of Huang et. al. [7] and Pavlakos et. al. [72], where term weights change during each stage. By performing the fitting in stages we are

Term \ Stage	1	2	3	4
λ_{data}	1	1	1	1
λ_{θ}	404	404	404	404
λ_{β}	200	200	200	200
λ_{a}	10	10	15	30
λ_{table}	50	50	50	60
λ_{to}	40	40	50	50
λ_{h}	0	0	10	20
λ_{sp}	0.00	0.00	0.01	1.0

Table 3.1: λ_{term} weights for $term \in data, \theta, \beta, a, table, to, h, sp$.

capable of guiding the optimization function such that certain terms become less or more important depending on the stage, the rationale behind such an approach is to first fix the "coarse" pose and then refine it. The term weights given each stage can be seen in Table 3.1. The term weights for the terms in Eq. 3.4 were found through manual adjustment to keep the terms balanced.

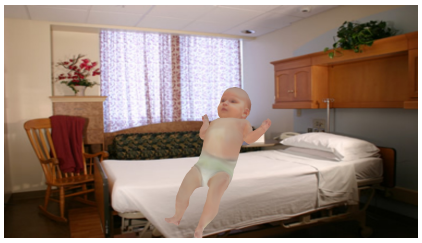
3.3.3 Post-Processing

Given the unconstrained nature of the 2D-to-3D lifting procedure, the fitting is likely to produce jittery results. Therefore we apply a post-processing step prior to the image generation. We apply a moving average filter of width 5 (2 frames in the past and 2 frames in the future) in order to smooth the parameters. For the first 2 and last 2 frames, we simply use the moving average of the available frames and do not perform padding. Doing so allows us to significantly smooth out pose and shape parameters between frames in a single sequence.

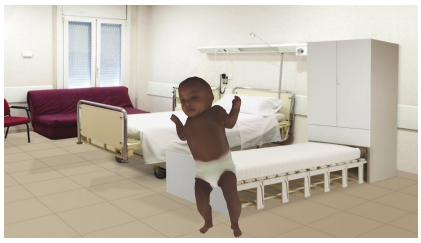
3.3.4 Video Generation

The video generation procedure uses the optimized registrations of the SMIL model to generate videos of synthetic infants based on real data. Similarly to how Huang et. al. [7] describe the process of creating a synthetic image, a synthetic video V_{syn} can be generated through an imaging process Ω as described in Eq. (3.10). Let R be an ordered sequence of optimized registrations and $C(d, f, t)$ be the single camera used in the video, where d, f and t are the principal point, the focal length and the translation parameters respectively of the camera. Additionally, let Bg be a randomly selected background image of a hospital room from the MIT Indoor Scene dataset [73] and T be one randomly selected infant texture provided by MINI-RGBD [8] that is mapped to the SMIL output mesh $M(\beta_i, \theta_i, \gamma_i)$. The SMIL mesh takes as parameters the β_i, θ_i and γ_i of the corresponding optimized registration $i \in R$.

$$V_{syn} = \langle \Omega(M(\beta_i, \theta_i, \gamma_i), C(d, f, t), Bg, T) \mid i \in R \rangle \tag{3.10}$$



(a) Synthetic Caucasian infant



(b) Synthetic Black infant

Figure 3.1: Individual frames of two different augmented videos from the same sequence.

In order to increase data quantity, for each sequence we replace the global pose parameters with a random rotation on the x, y and z axis. The x rotations are in the range $[\frac{\pi}{2}, \pi]$ while the y and z rotations are in the range $[-\frac{\pi}{4}, \frac{\pi}{4}]$; these ranges ensure that the infant will never be facing away from the camera or in an implausible recording angle. Figure 3.1 demonstrates a frame from two different synthetic videos augmented from a real infant video. In the figure each image has a different texture, background and global rotation. Note that the augmentation is achieved by the re-projection of the 3D joints after a rotation. By rotating the sequence and re-projecting to the image plane, we effectively create a new mapping of a 3D pose in 2D. To get an understanding of the variability that this video generation procedure is capable of, refer to Appendix A.

3.3.5 Occlusion Labels

Given that we are able to generate ground truth data for infant poses in video, we further extend this and generate occlusion labels for said poses. The occlusion labels generated follow the COCO format explained in Section 4.2, and are as follows: if the joint can be seen it is labeled as visible, if the joint is in the image but it is occluded (either by self-occlusion or by foreground occlusions) it is marked as occluded, and if the joint is not in the image it is marked as not visible. In order to generate the occlusion labels for the generated sequences we apply the Boxed Man model detailed by Wang et. al. [58] with a few adjustments.

Boxed-Man model

The Boxed-Man model is a planar model that describes the body segments, such as the head, torso, legs, and arms, as rectangles. Given a key-point and a set of rectangles, we can verify whether the joint is occluded by identifying whether such key-point is inside of another rectangle that it is not a part of. If a key-point is inside a rectangle, we compare the depth of the key-point to

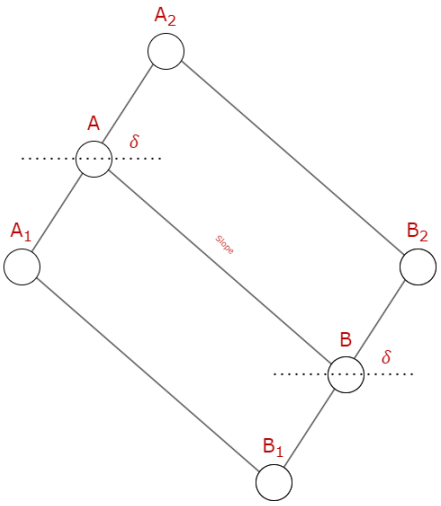
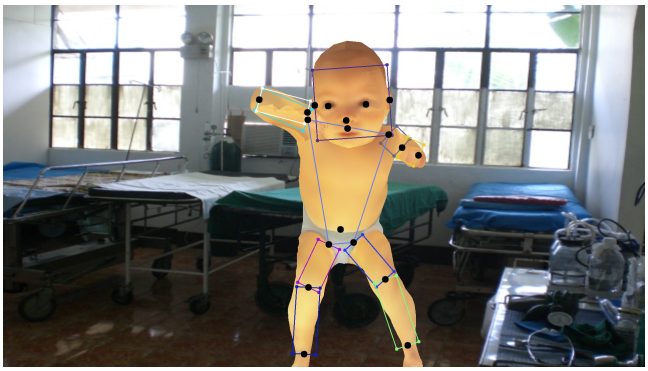


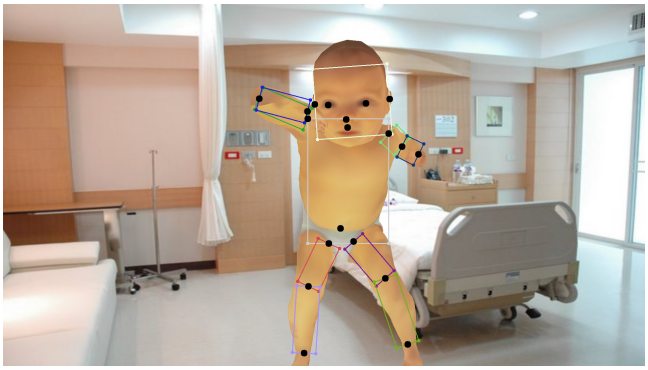
Figure 3.2: Box calculation supplementary help for the Boxed Man model. As indicated in the text, the two original points are A and B, additionally the slope between these points can be seen in the image. The distance between the generated points and the original point is determined by δ. Inspired by Wang et. al. [58].

the depth of the closest key-point in the rectangle. Each segment is originally composed of 2 key points, A and B, from which we can project into 4 points A_1, A_2, B_1 and B_2 to create the bounds of the boxes for each segment. The points are calculated by taking the perpendicular slope m_{AB} of the line AB, which is defined as the negative inverse of the slope $\frac{\Delta_y}{\Delta_x}$, applying a rotation, subtracting/adding the original key-point coordinates, and multiplying it by a scaling parameter δ. For specific implementation details, refer to Section 4.2 of Wang et. al. [58]. Figure 3.2 provides a visual aid for the box calculations. The results of such a method allows us to create boxes that are proportional to distance between two key-points in the image plane. Figure 3.3 shows an infant whose arms produce different box proportions - the left arm shows smaller box sizes, while the right arm shows proportionally larger box sizes. The black points on the infant body are the ground truth key-point locations.

Given that the model was originally developed with the Human3.6M skeleton in mind, some modifications had to be made. Rather than using the top of the head key-point and the thorax key-point for the head, we use the ear key-points of COCO19 in order to define the head box. For the head box we apply a δ of 105 and for the limbs we apply a δ of 35 in order to capture the different proportions of the infant. Furthermore, given that the hip-joints of the SMIL model are closer to the center of the hip, we use the maximum and minimum x and y coordinates of the two shoulders and two hips to create a



(a) Boxed-Man model prior to torso modification



(b) Boxed-Man model after to torso modification

Figure 3.3: The same frame of a synthetic infant with the original Boxed-man model and the modified Boxed-man model overlayed.

more rectangular shape; not doing so would result in incorrect occlusion labels. Figure 3.3 demonstrates the Boxed-Man model applied to the synthetic infant along side the corrected torso segment.

3.4 Results

In order to verify results a qualitative and a quantitative validation were performed. The qualitative validation was done manually, where we inspected whether a fitted sequence had severe mesh deformation (i.e. a crumbled up infant mesh). Any sequences which had such deformations, were removed from the dataset and from the subsequent quantitative evaluation. From the annotation data, the fitted sequences from videos 1 and 2 from Infant 2 were removed due to severe mesh deformations. In video 1, the mesh had severe inter-penetrations, the arms went through the head. In video 2, the head of the infant was fully

Infant	Video	WMSE	PCK@0.2
1	1	8.11	96.26
1	2	8.11	95.46
2	3	5.75	100
2	4	5.17	97.77
3	1	8.99	99.07
3	2	5.37	97.81
4	1	2.15	100
5	1	3.28	100
5	2	6.21	99.24
5	3	3.32	100
6	1	5.49	100
7	1	2.75	100

Table 3.2: WMSE in pixels and PCK@0.2 per sequence. Low MSE error and high PCK value indicate a good fit.

turned 180 degrees.

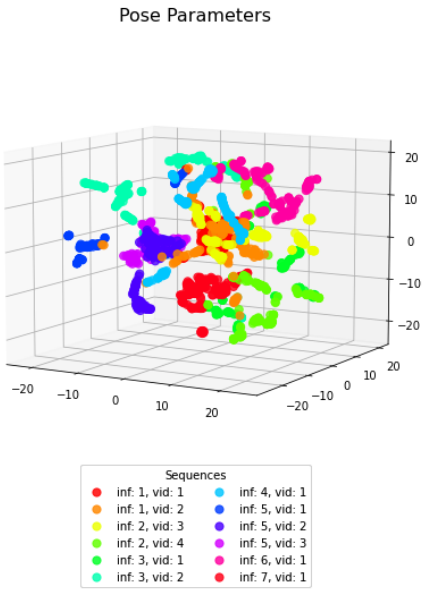
Lifting a pose from 2D to 3D is a highly unconstrained problem, and using a fitting procedure that minimizes an energy provides no guarantee that the pose is valid and correct. Given that there is no ground truth 3D data to evaluate the quality and correctness of the image generation we resort to using the 2D data. As SMPLify is a lifting procedure that takes into consideration 2D re-projection error, standard validation metrics such as PCK (Percentage of Correct Key-points) are not fully indicative of the quality of the results alone as they often provide very high scores. Therefore, the metric that we select to further validate the fitted sequences is the weighted MSE (Mean Standard Error), which is calculated in pixels. The MSE for a sequence is calculated between the re-projected 2D SMPLify joints and the ground truth annotations. Using the occlusion labels from the annotation, we set the weights of the key-points which are marked as occluded to 0 as they were not used during the fitting procedure and set the remaining weights to 1. The reasoning behind using this metric is that it allow us to measure how close the joints are in the image plane. It is important to note that given the fact only 2D is available, there is no measure that will effectively capture the quality of the results. The weighted MSE serves as an indicative, yet not infallible, metric for pose validity. Table 3.2 displays the PCK@0.2 and MSE for each fitted sequence.

T-SNE (T-distributed Stochastic Neighbor Embedding) was applied to the pose and shape coefficients of the SMIL model in order to understand the coefficient distributions of the sequences. Given that each pose is represented by 69 coefficients, we aim to reduce the dimensionality of our data. First we normalize the pose coefficients of the sequences by removing the mean and scaling to unit variance, to which we apply PCA. We reduce the data to 3 principal components

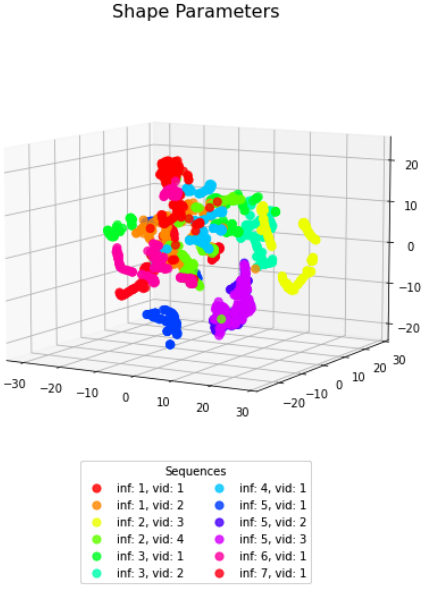
and feed it to the T-SNE algorithm. We optimize the T-SNE with a learning rate of 200, for 1000 iterations and apply early stopping if no change occurs for 200 iterations. Lastly, we use a perplexity of 15 due to the small sample size. We apply the same procedure to the shape coefficients. The results can be seen in Figure 3.4.

Figure 3.4a shows the shared pose coefficient clusters by sequence. This figure indicates that there are no severe pose coefficient outliers between sequences, all the sequences to some extend, share the same pose parameter space. Furthermore, this indicates that our pose variability for the registered data is good, with different sequences encompassing a large area of the parameter space. This provides further indication that the fitting procedure is capable of capturing a good range of motion and we do not get low movement sequences due to a severely constrained fitting procedure. Additionally a large portion of sequences have a big overlap between pose coefficients as seen in the figure around the (0,0,0) coordinates, indicating that we are able to capture similar infant positions across sequences. Having more fitted sequences would allow us to further collect more information about the infant motion range and how well our fitting procedure captures said motion. By inspecting the in-between clusters, we can see that most frames from within a sequence remain close to their respective sequence cluster. Figure 3.4b shows the share shape coefficient clusters by sequence. This figure shows that the shape coefficients have some variation within sequences. This is not particularly surprising as pose and shape coefficients are not independent from each other; certain poses affect the shape of the body. For example, slouching might make a person's torso appear thicker than it is. However, sequences Inf:1-Vid:1, Inf:2-Vid:4, and Inf:3-Vid:1 appear to have more variation than the other sequences, which indicates that the shape of the infants for these sequences might not remain accurately portrayed throughout the entire sequence or that the movement in these sequences significantly affect the shape of the infant.

In order to further visualize the results of the fitting procedure, we additionally plot the embeddings per infant. This can be seen in Figure 3.5, this is done to get an understanding of the overall parameter space and how much is shared between different infants. Figure 3.5a indicates that the pose clusters per infant are close in the embedding space, this indicates that we are able to capture the movements of the same infant across different videos effectively. However, this figure indicates that pose parameters across different infants might be significantly different. Given that our data relies on 2D points to perform 3D pose estimation this is not a surprising result and might indicate that the fitting procedure should be more constrained; future work should explore the use of VIBE in order to verify whether its a shortcoming of using a single image registration approach, or a shortcoming of these techniques. It is also possible that different infants move significantly different and strike different poses, this would explain the similarities in sequences from the same infant and dissimilarities in sequences between different infants. Figure 3.5b shows the shape parameters



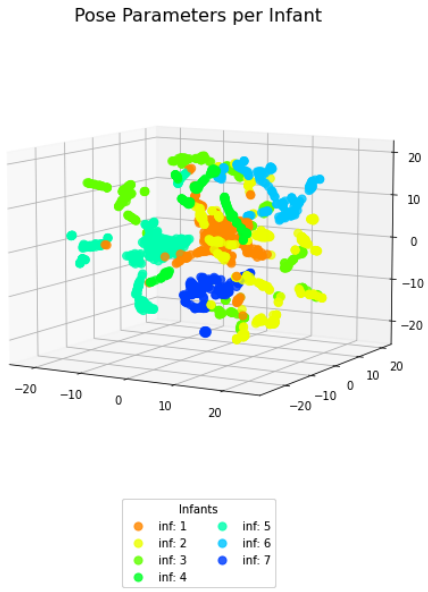
(a) T-SNE of the infant pose coefficients for all registered frames per sequence.



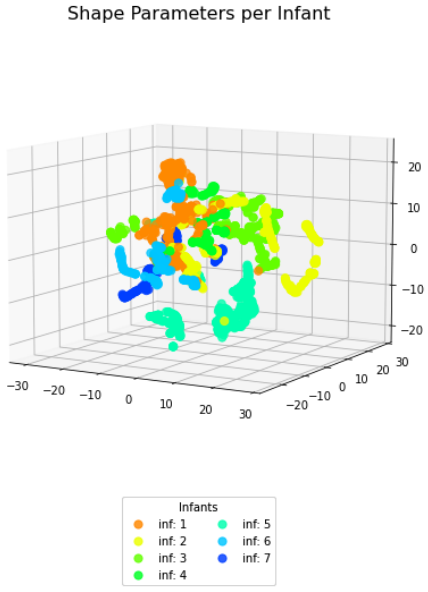
(b) T-SNE of the infant shape coefficients for all registered frames per sequence.

Figure 3.4: T-SNE plots for the pose and shape coefficients per sequence. The legend labels indicate the infant ID and its corresponding video. The results of 12 sequences are displayed. Each point in the plot represents a single frame.

per infant. This figure shows a even though the shape parameter space within infant sequences varies, the shapes between infants are different. This suggests that we are able to capture the different shapes caused by movements performed by infants. Additionally, this figure suggests that we are able to capture how the shape of different infants. This can be extrapolated from the fact that the sequences fitted from the same infant always produce shapes coefficients that are close in the shape space. These results might suggest that different infants do in fact move differently.



(a) T-SNE of the infant pose coefficients for all registered frames per infant



(b) T-SNE of the infant shape coefficients for all registered frames per infant

Figure 3.5: T-SNE plots for the pose and shape coefficients of sequence. The legend labels indicate the infant ID. The results of the 7 fitted infants are displayed. Each point in the plot represents a single frame.

Chapter 4

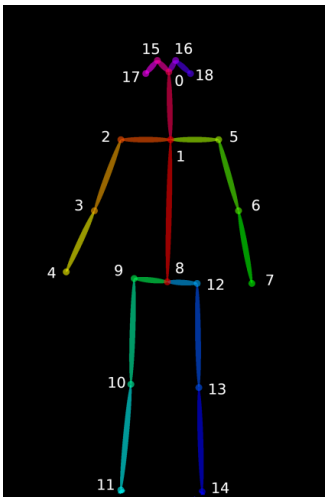
SPIS Dataset

4.1 Data Collection

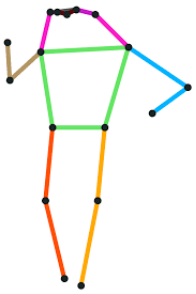
The data that was collected comes from preterm infants in the NICU at the University Medical Center Utrecht(UMCU). The data was recorded using a RealSense2 depth camera, however due to lighting limitations which prevent depth capture inside the NICU, only the RGB data was used. Each recording is stored as a 1920x1080 video and captured at an average of 30 frames per second. Due to the hectic nature of the NICU, the camera is placed in different locations between recording sessions. At its furthest, the camera was placed approximately at 50cm away from an infant and at its closest, at approximately 30cm. Each recording contains a single infant inside of an incubator with some form of foreground occlusion present. The original data contains hours of recorded and un-annotated footage, however for the purpose of this research, short clips of 5 seconds are annotated and used. These short clips were manually selected based on a set of criteria: lighting conditions, viewing angle, infant visibility and movement. The criteria of lighting conditions is simply based on whether or not the infant can be seen in the video. If the infant can be clearly seen, the video clip fits the criteria. For viewing angle, only video clips which did not have severe perspective distortion fulfill the criteria. Additionally, clips in which the infant displayed no motion were discarded. The reasoning behind using motion as a criteria is to gain as much pose variability as possible given the scarcity of the data in this domain.

4.2 Data Annotation

The collected data was annotated by three interns at the UMCU. The data annotated followed the COCO19 key-point standard (Figure 4.1a), as it was an appropriate annotation format for the fitting procedure outlined in Section 3.2 and could be easily retrofitted into the COCO17 key-point standard. Additionally, each key-point is annotated with an occlusion label. Each key-point



(a) COCO human 19 key-point standard. Retrieved from Cao. et. al. [46]



(b) COCO human 17 key-point standard. Retrieved from Kocabas. et. al. [74]

Figure 4.1: COCO19 and COCO17 Data annotation formats.

can be labeled as visible, partially occluded or occluded. If the key-point is not self-occluded or occluded by a foreground object the key-point is visible. If the key-point is partially visible, or occluded by the subject or by a foreground object, the it is labeled as partially occluded. Lastly, if the point is not present in the image, it is labeled as not visible. As previously mentioned, each clip is 5 seconds long and captured at 30 frames per second, resulting in a total of 150 frames per clip. There are in total 14 clips annotated from 7 different infants, resulting in approximately 2100 key-point annotated infant poses.

4.3 Data Pre-processing

As previously mentioned the data is annotated by hand, given the small movements that occur between frames annotators usually use previous annotations and make small corrections after a certain number of frames. Therefore we apply a pre-processing step to the annotations. In order to smooth out the annotations, we apply a moving average filter over the sequences with a width of 5, using two previous and 2 future frames. Lastly, given that the torso and limbs of the infant are always present in video, we relabel every joint above the hip from not visible to partially occluded; this is done in order to fix a set of key-point mistakes in certain annotations.

4.4 SPIS

For this research we introduce a hybrid dataset for preterm infants. The Synthetic (and Real) Infant Pose Sequences (SPIS) dataset was created in order to train and validate DL methods on preterm infant pose estimation in the NICU. This dataset is composed of a combination of real infants from the NICU and generated synthetic infants from these sequences. Each sequence is treated as its own independent data entry.

Like SyRIP, this is a hybrid dataset of real infants and synthetic infants generated with SMIL. Although the datasets use the same volumetric model and similar methodology, there are some key differences. SyRIP is an image dataset of full-term infants in-the-wild; the real data comes from publicly available video footage of infants. SPIS, on the other hand, is a video dataset intended to be used for motion modeling. Additionally, the real and synthetic data are from preterm infants in the medical domain. Lastly, the SyRIP dataset is approximately composed of 1000 synthetic and 200 real samples, while SPIS is composed of approximately 2100 real samples and 17700 synthetic samples.

In order to create a sizeable dataset that can be used to train and validate infant pose estimation, we augment each fitted sequence 10 times. Every augmented sequence has a different infant texture, a different background image, and a different viewing angle, which allows us to increase the visual variability of the data. Using the infant textures from MINI-RGBD [8], we are able to generate up to 12 different textured infants. Furthermore, there are 68 background images, which are hospital rooms selected from MIT Indoor Scene dataset [73]. Note that the hospital room images selected from this dataset have been manually filtered to not contain any people. The dataset is in COCO17 format, as seen in Figure 4.1b, which simply requires the removal of two points from the COCO19 format used to annotate the real data. The main difference between these two annotation formats, is that COCO17 does not contain the annotations for the thorax and the center of the hip. Furthermore, each sequence comes with key-point occlusion labeling generated by leveraging a modified Boxed-

Type	IDs
Real	3, 4, 5, 7, 8, 9, 10, 11, 12, 13
Synthetic	19, 25, 30, 40, 44, 50, 57, 64, 70, 82, 102, 106, 109, 118, 122, 132

Table 4.1: Real and synthetic sequence IDs in the SPIS testing set. The remaining sequence are in the SPIS training set.

Man Model [58] and foreground-background masks for future use. In total the dataset contains approximately 19800 frames of annotated and ground truth synthetic data.

The SPIS dataset has a total of 134 sequences, of which 14 real hand-annotated sequences and 120 are synthetic. We opted to use an 80%-20% split for training and testing respectively. Given that we want to validate the pose estimators for in-the-wild use, we create the test set to be composed of 10 randomly selected real sequences and 16 synthetic sequences. Sequence IDs for the real data range from 1-14, while 15-134 are for the synthetic sequences. The randomly selected sequence ids for testing can be seen in Table 4.1. The remaining data is used for training. Given the limited size of the dataset, no validation set is created, and we recommend to carefully use the training data for evaluation during training. The training split is primarily synthetic, making up 96% of the training data, the remaining 4% are real infant videos. Meanwhile, the test set has a more balanced real and synthetic data distribution, made up of approximately 38% real data, and 62 % synthetic data. Due to the scarcity of real data, we opted to use as little as possible for training; we found one third of the real data to be an acceptable compromise given the large amount of synthetic data available. For the test set, we selected enough synthetic sequences such that we would reach the required 20% split previously outlined.

Chapter 5

Preterm Infant Pose Estimation

5.1 Pipeline Architecture

This architecture is a single person, top-down, multi-stage approach. The architecture assumes that it will be fed both a series of images with infants present and their corresponding bounding boxes. The pipeline architecture is relatively simple: it contains a 2D pose estimator for encoding image features and decoding them into heat-maps that represent the location of joints, by selecting the peak activation of the heat-maps we find the predicted location of the joints in an image. Given a series of these predictions, a causal 2D TCN model is applied to predict the latest pose given the entire sequence of poses up until the current point in time.

5.1.1 Design Choices and Assumptions

Top-down

A shortcoming of top-down architectures is that they tend to suffer from early commitment issues. However given the nature of the domain and intended use, we believe this is appropriate. There are two main reasons as to why this is the case. Given the domain and future use, we can assume that pre-term infant pose estimation will be done with a single infant in mind. Due to the nature of the domain, preterm infants will always be alone and placed inside of the same environment: an incubator. Top-down architectures require a person detector in order to find the person in an image. However, given the nature of our data and the provided annotations previously mentioned, we train the model using ground truth bounding boxes. Future research should train a detection model, such as Mask-RCNN, to detect incubators or occluded infants for in the wild use. In this research we assume the ground-truth detection boxes are present.

Multi-stage

Often multi-stage approaches allow us to effectively control models more effectively. The reasoning behind using a multi-stage approach is due to the choice of using the FiDIP training framework for the 2D pose estimator and the format limitations provided by using the COCO annotation format. Originally we had the intent of using Human3.6M to train the TCN to perform 2D-to-3D lifting. However the available Human3.6M [75] data did not contain the key-points necessary to represent the poses in the COCO17. Therefore we opted for converting the 3D TCN into a 2D TCN that would be trained on the detections from a high performance 2D pose estimator pre-trained on COCO, such that the 2D TCN would be able to predict key-points in the COCO17 format of our data. This was done following the methodology of Pavllo et. al. [36], where they use pre-trained CPN and Mask-RCNN 2D point detectors and achieved similar results to those with ground-truth annotations. Producing key-points from our synthetic data in another format is not possible unless a joint-regression is learned between the SMPL model and Human3.6M meshed dataset, which we got denied access to; future work could address this limitation.

5.2 Pose Estimator

The architecture uses a 2D pose estimator in order to predict the key-points of the image. A wide variety of high performing 2D pose estimators for single frames exist, however very few of them have been tested or trained for infant pose detection. Therefore we select a model that has been evaluated on infant data previously, the DarkPose+FiDIP model, which uses an HRNet-W48 backbone, proposed by Huang et. al. [7] was selected as the underlying architecture to use for the 2D pose estimator. The reasoning for selecting this architecture as the foundation is two fold: (1) it has been previously tested in a similar domain and achieved exceptional results, (2) the training framework proposed by FiDIP embeds both synthetic and real data features in the same feature space. Using the FiDIP framework to train the 2D pose estimator is necessary in order to improve performance given our the data scarcity problem we face, and the need to use synthetic data to train our network.

5.2.1 Pre-training Procedure

The 2D pose estimator uses the FiDIP framework to transfer the knowledge of adult pose estimation to infant poses estimation by using a domain adaptation technique. FiDIP uses an adversarial approach to fine-tune a network with the main task of predicting poses and the auxiliary task of predicting whether the image is real or synthetic. In order to do so, they add an auxiliary domain classification sub-network, which takes as inputs the spatial feature representations of the pose estimator. For more in-depth information regarding the domain sub-network architecture, please refer to Huang et. al. [7].

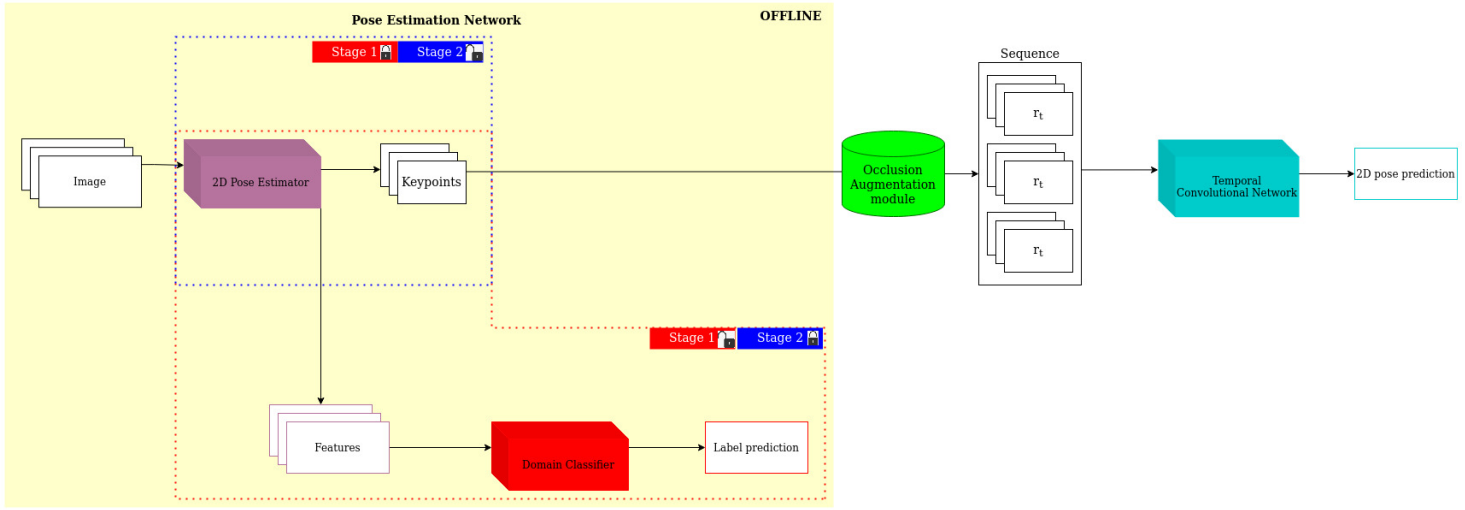


Figure 5.1: Pipeline Architecture. Here the blue and red rectangles with the locks represent the training stages of the pose estimator. The dotted outlines indicate which parts of the model are trained during their respective stage, indicated by the color. The locks represent whether or not we propagate the loss of the predictions to the networks at that stage. An open lock represents that we propagate the loss, while a closed lock represents frozen weights.

(a) Sample from SURREAL

(b) Sample from COCO

Figure 5.2: Image samples from SURREAL and COCO used to train the Domain sub-network.

Domain Sub-network

In order to effectively use adversarial training to encode the synthetic and real images in the same feature space, Huang et. al. [7] pre-train the domain classifier. Not doing so would degrade feature embedding performance and would render the use of the framework futile. The pre-training dataset is not publicly available, therefore we create our own. We randomly sample 2000 images of people from the COCO val2017 dataset and 2000 images of synthetic humans from the SURREAL dataset. Figure 5.2 contains an image from each dataset used to pre-train the sub-network. In order to ensure a diverse synthetic split, we randomly sample a single frame at a random time from 2000 different videos. We pre-train the classifier sub-network by propagating the Binary Cross Entropy (BCE) loss between the ground truth labels and the predicted labels. The sub-network is trained for 20 epochs, using AMSgrad as an optimizer, with a learning rate of 0.0001. The training data is augmented by randomly applying scaling with a factor of 0.35, rotations of at most 45 degrees and horizontal flipping. During the pre-training we use a batch size of 8.

Darkpose + FiDIP

Using the pre-trained weights of the HRNet-W48 and the pre-trained weights of the domain sub-network from the previous section, we train the Darkpose + FiDIP model on SyRIP to serve as our baseline for 2D infant pose estimation. We use SyRIP whole set to train the network. SyRIP whole set consists of 1000 synthetic infant images and 200 real infant images. We leverage SyRIP Test100, a small validation set of real infant images, to validate the baseline results.

Following the methodology of Huang et. al. [7], we freeze the layers of the HRNet up to the second stage and its bottleneck. We tune the network for 20 epochs, using a batch 8, AMSgrad as an optimizer, with a learning rate of 0.0001. Training data augmentation is performed by randomly applying scaling

with a factor of 0.35, rotations of at most 45 degrees and horizontal flipping. Similarly to the results outlined in the original research, the model achieves a 93.9 mAP, a 98.4 AP50 and a 98.4 AP75 on SyRIP Test100.

5.2.2 Fine-Tuning Procedure

Following methodology from Huang et. al. [7] we freeze the layers of the HR-Net up to the second stage and its bottleneck. By doing so, we are able to use the pre-trained weights to capture the early spatial feature representations and fine-tune a smaller set of parameters for our preterm infant pose estimation network using FiDIP; this is done to perform transfer learning given our small data domain. Given that our model has previously been fine-tuned and adapted to the infant domain using SyRIP, but not to the preterm infant domain, we train the same layers in hope of capturing new information about the preterm infant domain. The fine-tuning of the network is done for 5 epochs with a batch size of 16, using AMSgrad as an optimizer, with a learning rate of 0.0001. The reason for tuning the network for a few epochs lies on the fact that the synthetic data of SyRIP is also created using SMIL, therefore we aim to not over-fit to the synthetic data given our mostly synthetic training split. We fine-tune and test the network using SPIS, following the data split provided in Section 4.4. Training data augmentation is performed by randomly applying scaling with a factor of 0.35, rotations of at most 45 degrees and horizontal flipping to the images and key-points.

For every iteration, the batch of input images is provided to the pose estimation network, from which we collect two outputs. First, we collect the resulting spatial feature representations of the network for each image in the batch, which we feed to the domain classifier to identify whether the images are real or synthetic. We calculate the BCE loss between the ground truth labels and the predicted labels and propagate it through the domain classifier sub-network. Next we calculate the weighted MSE loss between the predicted heat-maps and the ground truth heat-maps of the key-points. Given that our data is annotated using key-points, we generate the target heat-maps using a Gaussian distribution with a $\sigma = 3$, and use the (x,y) coordinates of the key-point as the mean. Given the MSE and BCE loss of the batch, as indicated by Huang et. al. [7], we calculate the loss of the pose network as the following:

$$loss = MSE - (\lambda * BCE) \tag{5.1}$$

where λ is set to 0.0005. Note that this loss is only propagated to the pose estimation network. By doing so, the FiDIP framework guides the network to not only predict the location of the joints, but to embed the extracted features into the same feature space using adversarial training.

5.3 2D TCN

Given that the aim of this paper is to create a pipeline for preterm infant pose estimation in video, a network to model motion was required. We considered the optimal choice for modeling sequences of movements of preterm infants would be TCNs, as previous work [36, 43, 60] have demonstrated promising results for 2D and 3D pose estimation on adults. TCNs use 1D convolutions in order to capture and model sequences. Given the nature of our task, we use a causal and dilated convolutions for our network.

Causal TCNs are models which produce an output at time t and only perform convolutions on elements prior to and up to time t, therefore no future information is used. These models are often preferred in real time systems where information should be processed as it arrives [36]. Additionally, the choice to use a dilated convolution architecture is due to the fact that simple causal convolutions only allow us to increase the receptive field of the network linearly. Creating a model with a large receptive field using only casual convolutions would heavily increase memory requirements and slow down training as certain outputs would have to be recalculated. Dilated convolutions allow us to increase the temporal receptive field of the network while maintaining a reasonable model size and prevent us from re-calculating intermediate results. Figure 5.3 demonstrates how causal and diluted convolution TCNs work.

Although previous works [60] have experimented with training networks to capture multi-scale temporal by using multiple temporal strides in order to deal with varying degrees of motion (in particular for very fast motions such as those associated to sports), we refrain from doing so. The reason being that the range of speeds at which pre-term infants are capable of moving is small and movement speed variation in videos is low.

5.3.1 Architecture

The developed architecture can be seen in Figure 5.4. The TCN takes as input a sequence of 27 frames, each containing 17 key-points in 2D. The input is given to 1D convolutional layer with a filter width of 3, dilation of 1, and 512 channels, followed by a 1D batch normalization layer, an activation layer and a dropout layer. Immediately after two temporal blocks are applied and lastly a 1D convolutional layer with a filter width of 1, dilation of 1 and 512 channels. The network produces as an output a single frame prediction of 17 joints in 2D.

Each temporal block is made up of two components. The first component is a 1D convolutional layer of 512 channels and a filter width of 3, followed by a batch normalization layer, a ReLU activation layer and a dropout layer. The second component follows with 1D convolutional layer with a dilation of 1 and a width of 1, followed once again a 1D batch normalization, activation layer, a drop out layer and lastly a residual is added to the output. The temporal blocks

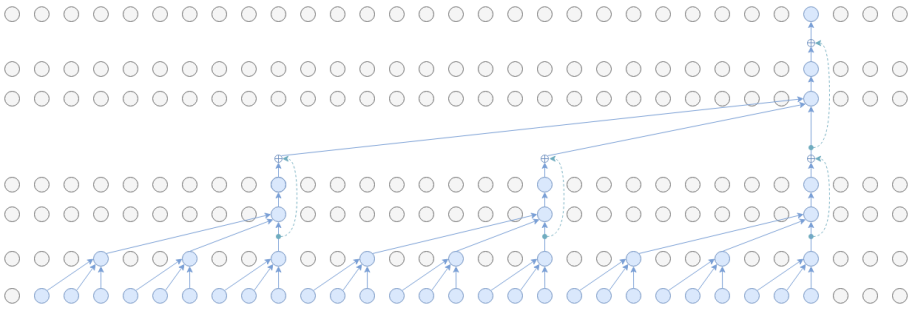


Figure 5.3: Causal and dilated convolutions of a generic TCN composed of 2 temporal blocks with filter widths = (3,3,3), stride = 1. In this example the TCN has a receptive field of 27 frames. The dilations have the same size as the filter widths and can be seen increasing the larger the receptive field becomes. Image retrieved from Pavllo et. al. [36]

.

can be seen in Figure 5.4 inside of the light orange rectangles. After the last temporal block, one last 1D convolution is performed to get the predicted key-points of a sequence. The blocks are created to have filter widths of size 3 accordingly. The dilation rate for each of these blocks is set to be equal to the width of the filter for each block. It is common to set the dilation rate equal to or less than that of the filter widths as making them larger would result in information loss. Setting the dilation rate of a block to be less than that of the filter width leads to reusing information present in previous filters and was purposefully not done.

The architecture was selected to have equal filter widths in order to keep the model relatively small while achieving a medium size temporal receptive field. The receptive field of the model can be calculated by multiplying the filter widths at each level. In our case the total receptive field of the model is 27 frames. With the first 1D convolutional layer having a receptive field of 1, the first temporal block a total receptive field of 3, and the second having a total receptive field of 9, followed by the final 1D convolutional having a total receptive field of 27 frames. Even though large temporal receptive fields often produce better results [36, 43, 60], there is a reason why we do not design our TCN to have a larger receptive field. The aforementioned research has been evaluated on adults performing action sequences and therefore motions often are highly correlated to one another. For example, it would be unlikely for an adult who is sitting down to to raise their arms in the air and move them back and forth. However, in the infant domain, motion is often less predictable and movement is less correlated. Second, our clips are relatively short (approximately 150 frames) and therefore a balance between sequence quantity and sequence length for training had to be taken into consideration.

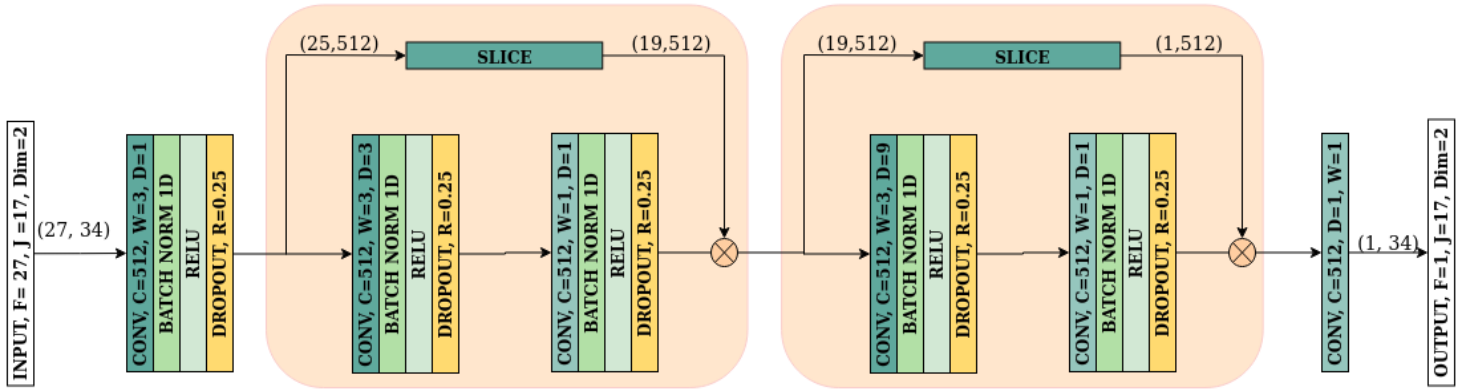


Figure 5.4: TCN Architecture as described in the text. The orange rectangles each represent a single temporal block. Here C represents channels, W represents filter widths, D represents dilution. Tensor sizes are also shown between parentheses, where e.g. (25,512) represents 25 frames and 512 channels.

5.3.2 Occlusion Augmentation Module

The occlusion augmentation module presented is inspired by the occlusion module proposed in Cheng. et. al. [60]. However some modifications are made due to architecture differences and lack of explicit information. In their implementation they apply the augmentation module used on heat-map outputs, while the current implementation works on key-point outputs. Additionally, while the goal of training TCNs to be robust to occlusions is similar, the methodology is significantly different. Given the preterm infant domain where there are significant occlusion, infant movements are short and actions are seemingly random, the occlusion module aims to simulate when no predictions can be made using a key-point detector. To do so, the module sets the x and y coordinates of the key-points to (-1,-1). Doing so, we hope to train network to identify that key-point which are set to (-1,-1) provide no temporal information regarding movement.

In order to do so, the module applies one of three occlusion types to the sequence. The three types of occlusions applied are joint occlusions, frame occlusions and location occlusions. All occlusions occur in a series of consecutive frames. Joint occlusions are occlusions that occur on a particular set of joints, regardless of location, in a sequence of frames. These are meant to replicate self-occlusions and foreground occlusions. Frame occlusions are those in which the entire body is occluded, and are meant to simulate heavy foreground occlusions and missed detections. Lastly, location occlusions are those which occur in a specific spatial location and are meant to represent persistent foreground occlusions, such as medical equipment attached to the infant.

Given that the proposed 2D TCN model has a receptive field of 27 frames, the number of consecutive occluded frames are sampled from a Gaussian distribution with a mean of 13, and a standard deviation of 1. Therefore, the occlusion module will apply an occlusion to at least 9 frames and at most 16 frames. The reason for creating such large occlusion windows is based on the fact that infant video is often heavily occluded. Due to the fact Cheng et. al. [60] provide no information as to how often their occlusions are applied, nor whether these occur on every batch or sequence, we follow conventional regularization methodology (such as dropout) and apply the simulated occlusions to 25% of the total batches.

5.3.3 Pre-training Procedure

Human3.6M

Human3.6M [75] is a MOCAP dataset that contains 8 actors performing 16 actions. The dataset uses 4 cameras in different locations in order to capture 3.6 million ground truth 2D poses of humans and their corresponding images. This dataset is originally intended for 3D pose estimation, however for the purpose

of our research we use the images for 2D pose detection. Furthermore, given that infant videos often do not have the infant parallel to the camera, we apply a random rotation augmentation from the set (-135, 90, 45, 45, 90 , 135) to the training key-points of each sequence of 27 frames; not doing so would lead to significantly worse results as indicated in Appendix B.1. If the rotations are not applied, the network only understands motion parallel to the ground given that Human3.6M is of adults performing actions.

2D Detection

The TCN is pre-trained on the Human3.6M dataset, however the key-points used to train the network are not the ground truth key-points of this dataset, but rather the detections of a Dark-pose model with a HRNet-W48 backbone pre-trained on the COCO dataset. Given that DarkPose is a top-down network and bounding boxes are required we use the ground truth bounding boxes of Human3.6M as outlined by Pavllo et. al. [36]. By training the network on the COCO detections rather than Human3.6M ground truth key-points we do not expect the network to learn a mapping of joints from Human3.6M to COCO. Given the small data domain we want to fine-tune the model to, expecting the network to learn a new mapping of joints is not feasible without severely over-fitting the network to the sequences. The detections of the network were collected and stored offline, and loaded in during training, which allowed us to speed up network training. Instead of using raw key-point detections, we normalize the key-points using the corresponding intrinsic camera parameters such that the screen coordinates are no longer from [0,max), but rather from [-1,1]. Normalizing screen coordinates allows us to use the model for videos of different resolutions. Following common methodology [36, 43, 60] we use the detections of subjects 1,3,5,6,7 and 8 for training. We guide the network by calculating the MSE error between the target and predicted joint locations.

Hyper-parameters

We pre-train the network for 60 epochs, using a batch size of 64 and apply random shuffling of sequences; sequences from different subjects are given in the same batch. Furthermore, we use an AMSgrad optimizer with an exponential learning rate decay. The initial learning rate is set to 0.001 and we apply a decay factor of 0.95. The drop-out rate is set to 0.25, and train with a temporal stride of 1. Additionally, for each epoch we decay the batch normalization momentum with the following equation:

$$m = m_{init} * \exp((\frac{p_i}{p_{end}} * \log(\frac{m_{init}}{m_{final}})) \tag{5.2}$$

where the m_{init} stands for initial momentum, m_{final} for final momentum, p_i for epoch i and p_{end} for the final epoch.

Pre-training Results

In order to verify whether the pre-training of the networks was done effectively, we leverage the COCO detections of subject 11 from Human3.6M. We use MSE (in pixels) and PCK@0.2 in order to quantify network performance. The results for the baseline 2D TCN and the Occlusion Augmented 2D TCN can be seen in Table 5.1 and Table 5.2, where total and per joint MSE and PCK@0.2 are displayed respectively.

Metric \ Condition	Baseline	Occlusion Aug.
Nose MSE	7.1	11.2
Eyes MSE	6.9	11.2
Ears MSE	7.2	11.5
Shoulders MSE	7.1	10.5
Elbow MSE	8.4	10.2
Wrist MSE	10.8	13.0
Hip MSE	6.3	7.9
Knee MSE	6.6	8.2
Ankle MSE	8.5	9.0
Total MSE	7.8	9.2

Table 5.1: Total MSE and per joint MSE in pixels for the baseline TCN model and the baseline occlusion augmented TCN on the H36M COCO detection test set.

Metric \ Condition	Baseline	Occlusion Aug.
Nose PCK	94.76	89.65
Eyes PCK	95.06	89.51
Ears PCK	94.17	88.79
Shoulders PCK	95.58	91.04
Elbow PCK	92.4	89.30
Wrist PCK	87.3	81.65
Hip PCK	96.02	94.64
Knee PCK	94.36	93.25
Ankle PCK	92.85	91.72
Total PCK	93.55	89.97

Table 5.2: Total PCK and per joint PCK scores for the baseline TCN model and the baseline occlusion augmented TCN on the H36M COCO detection test set.

The results indicate that both networks achieve acceptable results on the

validation set. The MSE for all joints is very low and the PCK@0.2 scores are high. This suggests that the 2D TCNs, both the pre-trained baseline and pre-trained occluded baseline, achieve good results on a large scale dataset and have been effectively pre-trained to the adult motion domain. These models are considered to be acceptable for fine tuning to the pre-term infant domain. However, given the lower PCK scores and higher MSE, we can infer that the augmentation module might not be leading the network towards using other joints to robustly predict joint locations. Given that the occlusion module effectively removes temporal information from the sequence, the network might simply be learning to only use the incomplete information. Rather than relying on other points for information, the network could be simply using the data available.

5.3.4 Fine-Tuning Procedure

Similarly to the methodology for the pose estimator, we fine tune and test the network using the SPIS dataset split proposed on Section 4. The reason for following the same split is to prevent artificially improving results by using previously seen detections. Given that we want to leverage the sequence information learned on Human3.6M, we only unfreeze the second temporal block of the network and the last 1D convolutional layer. The reason for doing so is as follows: sequence information at the first convolutional layer and the first block should not contain overall sequence motion given that they have a low temporal receptive fields, but are more likely to contain information regarding motion direction and velocity. We fine tune the network for 30 epochs, using the same hyper parameters as for the network pre-training.

Chapter 6

Results

6.1 Ablation Study

In order to study the impact the augmented training dataset, the occlusion augmentation, and evaluate the pipeline for preterm infant pose estimation we perform an ablation study. By studying the results of the network under 6 different conditions we are able to analyze which components improve preterm infant pose estimation in video. The conditions can be seen in Table 6.1, the main comparisons that will be drawn in the discussion of results will be between Conditions 1 and 2, 1 and 3, 1 and 4, 1 and 6, 3 and 4 and lastly 5 and 6. Conditions 1 and 2 use the 2D point detections of FiDIP, while Conditions 3-6 use the 2D point detections of the preterm infant tuned FiDIP model. We inspect and analyze the overall MSE error and PCK@0.2, as well as the per joint results in order to understand which joints seem to be the most difficult to predict. In the following subsections, the results will be presented; for an in-depth discussion of said results refer to Section 7.4.

Condition	Name	Fine-tune FiDIP	Occlusion Augmentation	Fine-tune TCN
1	TCN-B	No	No	No
2	TCN-OA	No	Yes	No
3	TCN-DA	Yes	No	No
4	TCN-DA-FT	Yes	No	Yes
5	TCN-DA-OA	Yes	Yes	No
6	TCN-DA-OA-FT	Yes	Yes	Yes

Table 6.1: Ablation Study Conditions. Here TCN stands to temporal convolutional network, B stands for baseline, OA stands for Occlusion Augmentation, AD stands for domain-adapted and FT stands for fine-tuned. AD indicates that the condition uses fine-tuned 2D pose estimation detections, while FT indicates that the TCN of the condition is fine-tuned to infant motion.

6.1.1 2D Pose Estimator Results

In order to study the impact of the synthetic data for infant pose estimation in images, we collect the results of the 2D pose estimator on the SPIS test set, which is composed of 10 real infants and 16 synthetic infants. To quantify their performance we use the mAP (mean Average Precision) as the main performance metric. The results for both the baseline FiDIP model and the tuned model for preterm infants can be seen in Table 6.2.

Model \ Metric	AP	AP50	AP75	AR	AR50	AR75
Baseline Model	53.6	76.0	60.3	55.6	78.0	61.5
Fine-tuned Model	82.1	98.0	87.8	83.8	98.7	90.0

Table 6.2: Model results for pose estimators on the complete SPIS test set. Table displays the Average Precision (AP) and Average Recall (AR) metrics. Additionally their variants AP50, AP75, AR50 and A75 are displayed, where AP75 and A75 are more strict metrics given the bounding box of the infant.

These results indicate that the re-training of the model aided the 2D pose estimation performance significantly. The baseline FiDIP model achieves a mAP of 53.6, while the fine-tuned model achieves a mAP of 82.1. Additionally, the mAR of the baseline FiDIP model achieves a 55.6, while fine-the tuned model achieves a 83.8. In order to further investigate the performance of the network, we test the network once again using only the real infant images of the SPIS dataset. By doing so, we are able to have a better understanding of the performance gain for in-the-wild use. The results can be seen in Table 6.3. The performance of the fine-tuned network significantly outperforms the baseline FiDIP model, in both mAP and mAR. The fine-tuned model achieves a 62.5 mAP on the real data, while the baseline model achieves a 15.9 mAP. In terms of mAR, the tuned model achieves a mAR of 69.7, while the baseline model achieves a value of 22.7. These results allows us to infer that the fine-tuned model has been adapted for the preterm infant domain.

Model \ Metric	AP	AP50	AP75	AR	AR50	AR75
Baseline Model	15.9	34.4	14.9	22.7	45.6	21.3
Fine-tuned Model	62.5	95.9	77.8	69.7	96.6	84.0

Table 6.3: Model results for pose estimators on the real preterm infant test split of SPIS. Table displays the Average Precision (AP) and Average Recall (AR) metrics.

6.1.2 TCN Per-joint Results

For each of the conditions we used the MSE error and the PCK@0.2 in order to quantify the results of the proposed architecture for preterm infant pose estimation. The total MSE and per-joint MSE results on the entire SPIS test set for each of the conditions can be seen in Table 6.4, while the total PCK score and per joint PCK scores can be seen in Table 6.5. By inspecting the tables, we can see that model from TCN-DA-OA-FT had the lowest total MSE for the SPIS test set split, while TCN-DA-FT had the highest PCK.

Metric \ Condition	TCN-B	TCN-OA	TCN-DA	TCN-DA-FT	TCN-DA-OA	TCN-DA-OA-FT
Nose MSE	89.5	70.2	62.8	51.3	49.4	49.1
Eyes MSE	86.3	69.1	60.7	48.8	52.9	47.6
Ears MSE	94.6	81.3	62.4	52.8	59.1	51.0
Shoulders MSE	89.6	77.9	44.8	37.3	40.5	37.1
Elbow MSE	109.0	90.6	57.6	50.6	48.0	46.6
Wrist MSE	123.9	110.7	101.9	81.8	80.5	71.1
Hip MSE	112.2	106.9	64.1	41.1	60.0	45.4
Knee MSE	123.2	111.95	130.5	58.3	106.6	59.6
Ankle MSE	125.2	119.0	130.1	64.8	115.1	71.1
Total MSE	106.9	94.4	80.4	54.3	69.1	53.4

Table 6.4: Total MSE and per-joint MSE in pixels on the complete SPIS test split for all ablation conditions.

Metric \ Condition	TCN-B	TCN-OA	TCN-DA	TCN-DA-FT	TCN-DA-OA	TCN-DA-OA-FT
Nose PCK	62.8	64.5	62.1	68.1	65.6	60.6
Eyes PCK	63.8	63.6	60.9	70.0	63.4	65.4
Ears PCK	41.4	35.9	42.2	53.6	36.4	52.4
Shoulders PCK	59.6	56.0	64.6	67.2	62.4	69.5
Elbow PCK	51.1	50.8	59.5	60.1	58.3	61.2
Wrist PCK	39.7	37.8	46.1	46.2	45.3	45.8
Hip PCK	23.3	19.8	49.4	68.9	55.9	65.72
Knee PCK	51.8	51.2	58.6	63.9	59.2	62.4
Ankle PCK	46.0	49.0	50.2	63.0	50.9	58.5
Total PCK	48.0	46.6	54.4	61.9	54.64	60.2

Table 6.5: Total PCK@0.2 and per-joint PCK@0.2 on the complete SPIS test split. for all ablation conditions. Note that it is abbreviated to PCK on the table.

Once again, in order to study their performance on only real preterm infant data and understand the differences between the synthetic and real data, we leverage the real data from the SPIS test split. Table 6.6 displays the total MSE and per joint MSE of the TCN models on only the real infant data. Table 6.7 displays the respective total PCK score and per-joint PCK scores. Again, the model that had the lowest MSE is TCN-DA-OA-FT and the model with the highest PCK score is the model from TCN-DA-FT. Additionally, it is clear that the MSE are significantly higher and PCK scores are significantly lower for all models when tested with only real data.

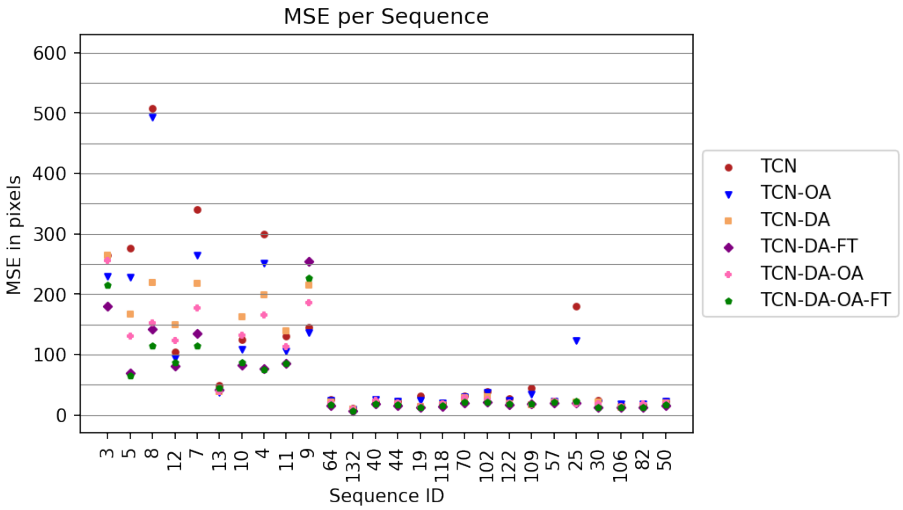
Metric \ Condition	TCN-B	TCN-OA	TCN-DA	TCN-DA-FT	TCN-DA-OA	TCN-DA-OA-FT
Nose MSE	168.5	132.5	138.9	111.3	102.6	102.7
Eyes MSE	161.5	130.3	133.4	104.9	111.8	99.2
Ears MSE	172.9	147.8	119.4	102.8	110.6	96.5
Shoulders MSE	180.2	155.1	88.1	71.0	74.1	69.7
Elbow MSE	231.2	192.0	119.6	105.0	94.7	94.0
Wrist MSE	245.9	222.4	219.4	176.2	165.1	146.5
Hip MSE	239.4	225.27	135.8	81.1	127.7	91.6
Knee MSE	293.9	264.54	318.9	136.5	257.5	140.3
Ankle MSE	290.7	274.5	309.2	148.8	268.7	163.8
Total MSE	223.6	197.4	178.1	115.5	148.4	112.1

Table 6.6: Total MSE and per-joint MSE in pixels on the real data SPIS test split for all ablation conditions.

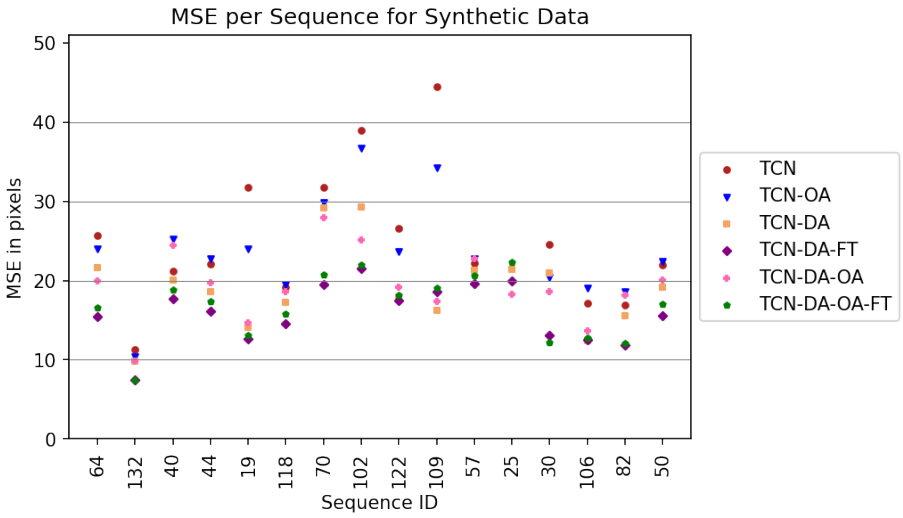
Metric \ Condition	TCN-B	TCN-OA	TCN-DA	TCN-DA-FT	TCN-DA-OA	TCN-DA-OA-FT
Nose PCK	36.2	42.6	26.5	38.2	39.2	24.8
Eyes PCK	35.9	40.5	24.6	40.2	31.4	31.8
Ears PCK	24.5	21.5	29.6	30.6	20.1	31.7
Shoulders PCK	38.9	42.8	41.3	49.1	48.9	53.0
Elbow PCK	22.9	24.8	31.9	28.8	31.6	33.0
Wrist PCK	12.7	9.9	14.1	11.6	15.1	15.3
Hip PCK	3.3	3.1	8.5	39.9	6.8	30.8
Knee PCK	6.4	7.7	6.1	17.9	6.2	13.6
Ankle PCK	1.9	2.7	0.9	18.6	1.9	8.7
Total PCK	19.3	20.5	20.0	30.1	21.3	27.1

Table 6.7: Total PCK@0.2 and per-joint PCK@0.2 on the real data SPIS test split. for all ablation conditions. Note that it is abbreviated to PCK on the table.

To get an understanding of the model performance per sequence, the total MSE per sequence were collected. Figure 6.1a shows distributions of the MSE each model for every sequence. These results provide us some insight regarding the overall model performance and the difference in performance between real and synthetic data. The same results are collected for the PCK scores of the model and are displayed in Figure 6.2a. The model performance is significantly different between the real and synthetic data, the synthetic data has significantly lower MSE overall and significantly higher PCK scores. Given these differences, we plot the performance of the models for only the synthetic sequences in order to gain a better understanding of the variation in performance. Figure 6.1b and Figure 6.2b show the zoomed in MSE and PCK of the models respectively. Note that the legends remain the consistent across figures. These results indicate that the real and synthetic data are significantly different; it is possible the synthetic data is not challenging enough for the pose estimator, creating easier to predict sequences.

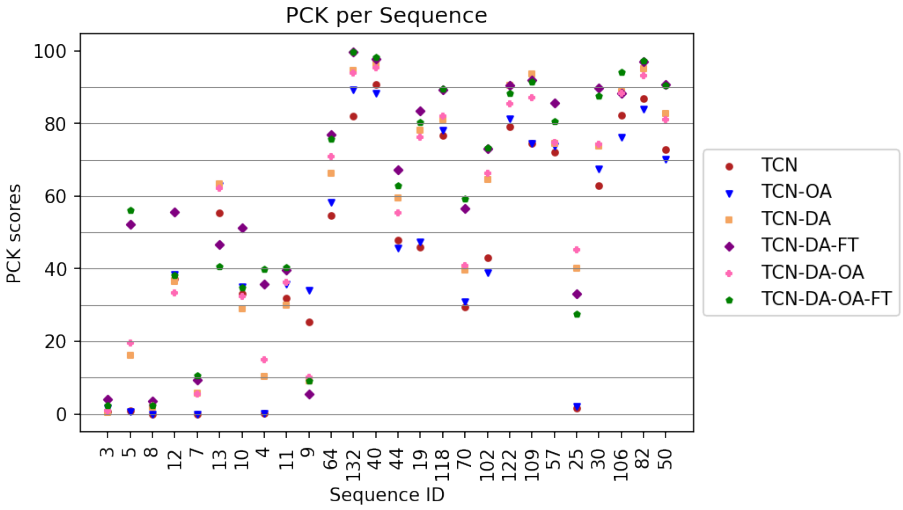


(a) MSE per sequence for the SPIS dataset.

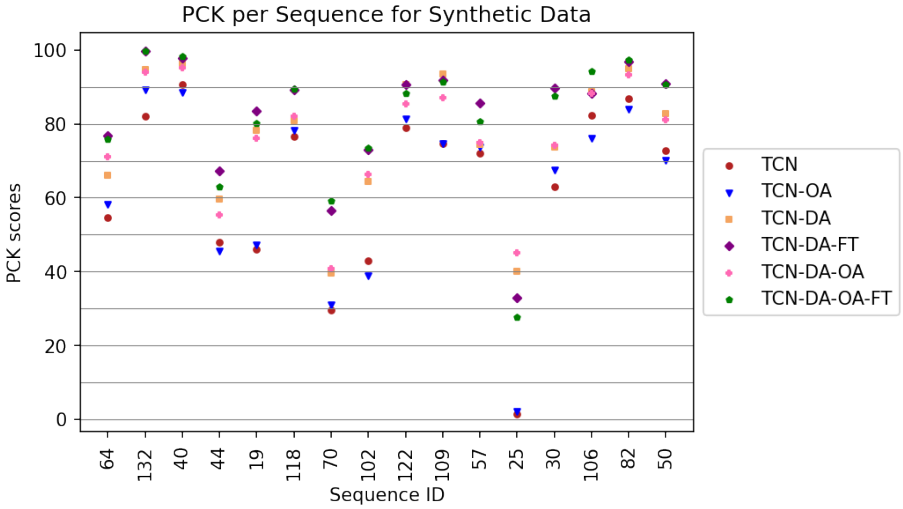


(b) MSE per sequence for only the synthetic data in the SPIS dataset.

Figure 6.1: MSE per sequence. The scatter plot displays the MSE of each model for every sequence. The x-axis is the sequence, while the y-axis is the MSE in pixels. The legend shows the corresponding label for each model, the legend labels remain consistent across sub-figures. Figure 6.1a shows the entire error distribution, while Figure 6.1b shows the error distribution over the synthetic data. Note the change in MSE range.



(a) PCK per sequence for the SPIS dataset.



(b) PCK per sequence for only the synthetic data in the SPIS dataset.

Figure 6.2: PCK per sequence. The scatter plot displays the MSE of each model for every sequence. The x-axis is the sequence, while the y-axis is the PCK score. The legend shows the corresponding label for each model, the legend labels remain consistent across sub-figures. Figure 6.2a shows the entire PCK distribution, while Figure 6.1b shows the PCK distribution over the synthetic data.

Chapter 7

Discussion

In the following sections we will discuss and analyze the results previously presented. First we will discuss the impact of using synthetic data on the performance of the TCNs. In order to do so we will mainly compare the results between the conditions TCN-B and TCN-DA. Next, we will analyze the impact of using the occlusion augmentation technique on the TCN performance. The main comparisons will be drawn between the conditions TCN-B and TCN-OA. Immediately after, we will study the impact of fine tuning the TCNs to infant motion. The main conditions that will be compared are TCN-DA and TCN-DA-FT. Lastly, we will evaluate the overall performance of the proposed pipeline by comparing the results from the conditions TCN-B and TCN-DA-OA-FT.

7.1 Synthetic Data

By analyzing the results of TCN-B and TCN-DA in conjunction with the results for the 2D pose estimator, we can identify that the use of synthetic data to augment pose and visual diversity for preterm infant pose estimation is effective. As previously stated in Section 6.1.1, the baseline pose estimator achieves a mAP of 53.6 and a mAR of 55.6, while the fine-tuned model achieves a mAP of 82.1 and a mAR of 83.8; seeing a 53% increase in mAP and a 50% increase in mAR. These are promising results and validate the use of the synthetic data for preterm infant pose estimation in images. This indicates that the FiDIP framework is effective not only for the infant pose domain, but can also be leveraged in the preterm infant pose domain. Additionally, the use of synthetic data generation using SMIL for the preterm infant domain is validated by the increased performance of the fine-tuned model on the real infant data, where the fine-tuned model significantly outperforms the baseline model in mAP and mAR due to the addition of largely augmented training data with a wide variety of viewing angles and visual variability. It is important to note that these results can only be validated for Caucasian infants, given that the real split of the testing does not contain infants from other demographics and ethnic groups.

However, we hope that with the use of diverse infant textures, results remain consistent.

Lastly, given the performance differences of the models on the SPIS test set and the real data, it is clear that the networks often perform better on the synthetic data. This can be particularly seen in Figures 6.1a and 6.2a. The MSE per sequence of the real data is often significantly higher for all models, and the PCK scores are significantly lower. Figure 6.2a indicates that from the synthetic sequences, number 25, 44, 70, 102 are particularly challenging; even when the MSE of these sequences are low. There are two possible explanations to the results that we see. One of the explanations might be that the synthetic data isn't sufficiently challenging for the pose estimator. The improved mAP and mAR of the fine-tuned pose estimation model seem to indicate this when comparing the results between test sets. Even though the results are significantly better on real data, the performance is not optimal. This supported by the fact that the the MSE of all models for all synthetic sequences is within the 0-50 pixel range, except for 2 outliers in Sequence 25. This is further supported by the difference in PCK scores of models on the synthetic data compared to the real data Another explanation might be found in the smoothness of movement in the synthetic data. Given that the annotation is performed by humans, even when smoothing is applied, certain discrepancies might arise. Additionally, the synthetic data might produce more predictable motion given that an additional smoothing is performed on the pose parameters in post-processing; since the points are smoothed during pre-processing and after the fitting in post-processing. For example, if the infant performs a fast and short twitch in the real data, the smoothing might remove or significantly decrease the movement seen in fitted sequences from said video. Furthermore, additional smoothness is also present in synthetic data given that there is no human error and the location of the joints on the body do not change per annotator, as synthetic data use vertex mapping to indicate joint locations and are precise.

7.2 Occlusion Augmentation

By comparing the results of TCN-B and TCN-OA we are able to identify that the occlusion augmentation module alone is not an effective module for preterm infant pose estimation. The MSE and PCK@0.2 of the baseline model performance sees a minimal increase when the occlusion module is introduced during training. The same can be said when we compare the results of the models from TCN-DA-FT and TCN-DA-OA-FT, where the MSE sees an insignificant decrease with occlusion augmentation, but the PCK value decreases as well. These results can be further visualized in Figure 6.1a, where the occlusion augmented models consistently show slightly less MSE compared their non-augmented counter parts on the real data split (Sequences between 3 - 13). By inspecting Figure 6.2a, we can see that PCK of the occluded models remains relatively unchanged when compared to their non-occluded counterparts on the

real data. Similar results can be seen when we compare TCN-DA against TCN-DA-OA. These results hold true for both the SPIS test set and the real data set.

Interestingly, the joints which are often occluded in the data, such as the ankles, hips, and shoulders obtain better detection and lower MSE across conditions which use occlusion augmentation during training; this can be seen in both test sets. There are several reasons as to why this might be the case. The TCN is dependent on the quality of the detections provided by the pose estimator used to predict joints locations, as seen by the results outlined in the previous section. The module removes temporal information from the sequence and, by design, attempts to force the network to identify motion through a sequence with incomplete information. Given that the network provides incomplete joints to the TCN, the TCN predicts the closest pose given the incomplete set of detections. If the detector produces very low quality results for a large quantity of the joints and further occlusion is applied, little to no information is given to the network about infant motion and it produces a result closest to the pose prior. Given that the baseline architecture already has difficulty predicting the spatial location and temporal movement of these joints, relying on the internal representation of the network for joint movements is not an effective solution for dealing with occlusions. Additionally, the occlusion module of the network applies occlusions randomly inside a sequence of frames, indicating that the information should be present at some time within the temporal window, however in the preterm infant domain occlusions are likely to be consistent across larger periods of time. Addressing this problem is difficult given that preterm infants do not perform highly correlated movement sequences and motion information from larger temporal windows likely will not help. Additionally, given the dependence of the pose estimator, training the TCN on the detections of the network might introduce additional noise. A solution to this issue would be to make the TCN and pose estimator trainable end to end rather than using a multi-staged approach.

Different ways of handling occlusion might provide better results, for example introducing the occlusion labels into the loss function of the network, as Chung et. al. [43] did for their 3D pose estimation network and leveraging a preterm infant pose prior might produce better results. Another potential solution would be to introduce an occlusion detection head to the 2D pose estimator, similar to the domain head of FiDIP. This occlusion detection head should be trained in order to predict whether or not a particular joint is present in the image but it is occluded by leveraging the occlusion labels. We would be able to implicitly model occlusion detection in images, similar to the work of [59], and leverage the occlusion maps to guide the subsequent TCN network to use said information. By using the Boxed Man model for in-the-wild videos and randomly masking images, generating enough ground truth occlusion labels to train such a module should be possible. Furthermore other occlusion augmentation techniques could be explored in conjunction with the previously discussed solutions, such as an augmentation module for synthetic occlusions

in the data generation procedure. Generating occluded data and leveraging it as ground truth to train an occlusion detection head for foreground occlusions could provide promising results. The generation of said occluded data was attempted during this research, however due to time constraints and the lack of portability of SMIL to rendering environments other than OpenDR [76], synthetic occlusions with realistic properties were difficult to generate and it was deemed unsuccessful.

7.3 Infant Motion

Given the preliminary results obtained in Table 5.3.3, it is clear that the 2D TCNs that were trained were able to accurately capture sequence information with minimal error. Therefore it is particularly surprising that the error of the baseline models and their subsequent fine-tuned variants have poor performance on the real preterm infant data. TCN-B demonstrates the reduced performance on the preterm infant data with baseline model detections, while TCN-DA demonstrates improved performance scores with fine-tuned pose detections. By comparing Condition TCN-DA and TCN-DA-FT, we can analyze how using the infant data to train the TCN impacts the performance of the TCN on infant motion. The results indicate that by fine-tuning the TCN, the performance of the model is significantly improved for both MSE and PCK, in the complete SPIS set and the only real data split. The results indicate that by using the synthetic data, the total MSE error decreased from 80.4 to 54.3, and the PCK score increased from 54.5 to 61.9. These results show a decrease of 32.5% in MSE and an increase of 13.8% in PCK score on the SPIS test split. For only real data, the MSE error decreased from 178.1 to 115.5, and the PCK score increased from 20.0 to 30.1. These results show a 35.1% decreased in MSE and a 50.5% increase in PCK scores.

Given that one of the performance bottlenecks of the architecture is the 2D pose estimator, it is crucial to have a pose estimator that performs well in order to feed the TCN with as much correct information as possible. The performance improvement for the pose estimator has a significant impact on the results of the TCN. As seen in Table 6.4, the MSE error for the majority of the joints significantly decreases for conditions TCN-DA, TCN-DA-OA, TCN-DA-FN and TCN-DA-OA-FT; these all use the fine-tuned pose estimator detections. Given the performance differences on the real and synthetic data as seen in Tables 6.4 and 6.5, as well as in Figures 6.1a and 6.2a, we can identify that the performance of the TCNs are limited by the 2D pose estimator performance. These figures indicate that performance of models on the real data has much higher variability, which indicate different degrees of detection difficulty. These have an impact on the infant motion results. The same cannot be said for the synthetic data, in which the pose estimator appears to perform extremely well and creates more complete and correct infant detections which improves the TCN performance

on these sequences. By looking at all these results, we can identify that in order to improve the performance of the model for infant motion, a better performing 2D pose estimator is needed.

The knee and ankle joints are the only joints for which the MSE is worse across these conditions even when using the synthetic data. There are two plausible explanations for this. These joints are often the most occluded in the infant domain, and especially in the real data, and their detection might not have been significantly improved in the fine-tuned pose estimator. This is made apparent by viewing the results of these joints in Table 6.6 and Table 6.7, which indicate low detection scores and high MSE. Therefore, sequences with extreme occlusions might have no information whatsoever regarding the positions of these joints. Additionally, the network was originally trained on Human3.6M, which is a dataset of adult poses. Due to the difference in body proportions, the information regarding motion might be modeled significantly differently and thereby impact the baseline results, this can be seen by the overall higher total MSE and lower total PCK scores when compared pre-training results of the TCNs shown in Tables 5.1 and 5.2. For example, the proportions of an infant could make the network predict a sitting or crouching motion for adults, and learned temporal information might produce erroneous results. By providing the TCN with infant data, the network improves at modeling proportions and movements of the infant more appropriately. These results, however, indicate that the fine-tuning of the network was not enough to obtain accurate estimations and positive results given the high MSE rates and low PCK scores across all conditions for which we used infant data to tune the TCN.

The PCK scores for conditions TCN-DA, TCN-DA-OA, TCN-DA-FN and TCN-DA-OA-FT indicate similar results with the majority of joints seeing improved detection scores when using the fine-tuned pose estimator detections, even for the ankle and knee. It is clear from the results of both data splits that the hardest joints to effectively capture are the knees and ankles. These joints have the highest error and lowest PCK scores across all conditions. Solely looking at the PCK for TCN-B AND TCN-DA does not provide a clear picture into the results. We can see that they have a negligible difference in total PCK score, but the per-joint PCK and per-joint MSE paint a much different picture. The MSE indicates that across sequences the error is lower, however, the PCK detects how often this error is within a particular threshold. This allows us to infer that even though the MSE is lower, the detection rate remains the same indicating that the performance gain is not significant enough on these joints. Additionally, given the lower MSE for the nose and eyes, as seen in Table 6.4 and Table 6.6 between TCN-B and TCN-DA, we can see that MSE as the sole metric does not help us to identify the best performing model. Even the error might be lower on average for TCN-DA when compared to TCN, it often is not enough to count as a detection.

Another reason for obtaining such results could be due to the fact that the

annotated data used to train the TCN and create the synthetic infants had inaccurate annotations in the knees and ankles, due to annotator error, which could not be fully pre-processed or removed from the data entirely. During the data generation procedure these key-points were simply marked as occluded and thereby removed from the fitting, which might have caused irregular motions in the knees and ankles, thereby reducing the reliability of these points for temporal modeling. This in particular would explain the high MSE rates present in the knee and ankle joints across all conditions, in particular with the real data split as seen in Table 6.6. Additionally, this would also explain the significantly higher total MSE and significantly lower PCK scores between the complete SPIS test set and real data results. For an analysis of the results using only the upper-body joints, refer to Appendix B.2. Even with the severe augmentation of preterm infant data, the resulting sequences only created enough data for 733 training sequences. This indicates that the results could additionally be impacted by the effects of data scarcity, which might be much more pronounced issue for motion data. A solution to this would be to first train the network from scratch using a dataset such as MINI-RGBD [8], which provides infant motion information and proportions. By doing so, we could freeze the weights once again and re-train a smaller number of network layers and fine-tune them for preterm infant motion. This was not considered in time during this research, future work should address it by leveraging MINI-RGBD and analyzing the performance of the model.

7.4 Pipeline Evaluation

Having analyzed the conditions and their respective results, we are now able to analyze the performance differences found between TCN-B and TCN-DA-OA-FT, and the overall architecture. TCN-DA-OA-FT outperforms TCN-B for infant pose estimation, with a MSE decrease of 50%. The model for TCN-DA-OA-FT achieves a lower total MSE and a higher total PCK for both the SPIS test set and the real data only. Indicating that the combination of fine-tuned detections and infant motion tuning outperforms the baseline network. Given that performance between conditions that differ only on whether or not occlusion augmentation is performed is minimal, we can safely infer that the most important factor is whether the TCN component of the pipeline is fine-tuned to the domain.

The results indicate worst performing variant of the pipeline is unsurprisingly the baseline model for TCN. This model has the highest total MSE across and lowest PCK scores across all joints, having a few exceptions with TCN-OA, in both the SPIS test set and the real data. This model attains a total MSE of 106.9 and a total PCK score of 48.0 on the SPIS data set, and a total MSE of 223.6 and a total PCK score of 19.3 for the real data set. By inspecting the results in all tables, and across every condition, it is clear that the best

performing model is the one for TCN-DA-FT with a total MSE of 148.4 and a total PCK score of 30.1 for the real data, and a total MSE of 54.3 and a total PCK score of 61.9 for the SPIS test set. Additionally, by inspecting the results in Figures 6.1a and 6.2a we can see that the results on the real data are significantly different between real and synthetic data. The most significant differences in performance arise from the real sequences, while the variations in performance of the models for the synthetic data remains small, the large differences occur in the real sequences. However, given the nature of the domain and goal of the task, every condition severely under-performed. There is ample room for improvement and the current results cannot be used trivially for downstream (or subsequent) tasks such as classification of the behaviors. Future improvements for the shortcomings of the components have been provided and future work seek to improve and iterate on this architecture.

Chapter 8

Conclusions

At the beginning of this research, we proposed a main research question, which was broken down into two different sub-questions in order to analyze and understand whether the methodologies were suitable for preterm infant pose estimation. The first sub-question we sought to answer was aimed at tackling the data scarcity of the preterm infant domain and was as follows:

Can a preterm infants movement be modeled in order to learn its pose parameters?

The application of SMPLify, in conjunction with the use of SMIL as proposed by Huang et. al. [7] were proposed in order to tackle this question. Two main conclusions can be drawn from the results. First, given that SMIL is a volumetric model of a full-term infant and not a premature infant, the increased performance of the 2D pose estimator allows us to conclude that physical differences between pre-term and full-term infants for infant pose estimation are not significant enough to require the use of a specialized volumetric model. This demonstrates that using SMIL for preterm infant synthetic data generation can be an effective method for tackling the data scarcity domain for images. However, given the lower performance score of the baseline FiDIP model, we can extrapolate that there are some significant differences between the preterm infant domain and the full-term infant domain, particularly in the viewing angle and observed poses. It is important to note that, in the work of Huang et. al. [7], additional textures (in particular, Adult SMPL textures) and body meshes were used to generate more diverse synthetic infants. This difference in performance could, in part, be attributed to that and future research should refrain from generating such synthetic data for medical applications.

In conjunction with obtained results and their subsequent analysis, it is clear that the performance of the TCNs on preterm infant data using augmented synthetic sequences is not yet satisfactory. The modified approach proposed in order to generate synthetic sequences does not provide the necessary qualitative results for infant motion modeling. This is clearly outlined by the results

obtained and the lack of performance gain obtained from leveraging said data during the training for motion modeling. The proposed methodology is effective for generating synthetic images of infants, but is not effective for modeling motion in videos. Further research into more sophisticated methodologies is necessary; some interesting avenues for doing so are proposed in the following section.

The second sub-question we sought to answer was aimed at addressing the challenge of occlusions in the preterm infant domain, and how to tackle these occlusions in video. The sub-question was as follows:

Can occlusion augmentation techniques used during training aid to minimize the errors for preterm infant pose estimation?

As indicated by the results previously discussed, the application of the proposed occlusion augmentation module was not satisfactory. Given the low performance variation between networks which use the augmentation module during the training, we deemed the technique to not be suitable for minimizing errors for the infant pose domain. However, further research could explore the avenue of using such technique for network regularization. Additionally, given the shortcomings of this methodology, an attempt to implement this module for heat-map representations might be fruitful. Using the heat-maps for temporal information rather than sole key-points might produce more promising results given that heat-maps tend to contain more information than sole key-points. The occlusion augmentation module designed was originally intended for heat-map representations, however given the challenges encountered during the pre-processing and training of the TCNs this was not possible.

Lastly, given that results and the answers of the previous sub-questions answered, it is now possible to answer the main research question which is as follows:

Can preterm infant poses be accurately estimated under occlusions in controlled NICU environments?

Given shortcomings in the data generation procedure for capturing infant motion and the lack of improvement with the occlusion augmentation techniques, in conjunction with the quality of the real data, and the design of the architecture the answer to this question with the current methodology is "no". As previously mentioned, the synthetic data is helpful for training the 2D pose network to detect poses in images, however the quality of the annotations and the resulting movements present in the sequences do not produce data that effectively captures the movement on infants in occluded sequences, which are produced when the 2D pose estimator produces low confidence results. The achieved mAP and mAR of the fine-tuned estimator demonstrate that the poses can be somewhat accurately estimated in images. However, given the target domain, higher AP scores are desirable. The same can be said for video, given the low PCK@0.2 scores and high MSE of all networks across all conditions.

Given that the preterm infant pose estimation should be accurate in order for it to be valuable for motion analysis, motion tracking and infant monitoring, the proposed architecture does not currently meet the requirements to provide a positive result to this research question. A positive outcome of this the research is that it has enabled us to understand that pre-training TCN networks on adult data and fine-tuning them to the preterm infant domain might not be effective strategy for tackling the data scarcity problem alone. As previously alluded to, the differences in body proportions and performed movements might already be represented in early layers of the TCN; this demonstrates that the TCNs should preferably be trained purely on infant data. By freezing the early layers of the TCN model, we assumed these only contained motion and direction information about joints, however it appears to be that body proportions and joint locations in the image might be represented in these early features as well. Given that the pre-training results and the collected results produce significantly different total MSE and PCK values, this might be the case. Lastly, we have learned that preterm and full-term infant physical differences are not significant enough to require a different template mesh and shape prior.

8.1 Future Work

The current methodology for generating synthetic data has a series of limitations that should be addressed in future work. These limitations can be divided into two different categories; limitations that arise from methodological choices and limitations that arise from the data domain. SMPLify was originally developed for single frames, yet our data comes from video footage, due to this spatio-temporal information cannot be fully captured using this method. To address this, future work would should follow the methodology proposed by Kocabas et. al. [77] and extend it by training a motion discriminator for infant motion levering a 3D infant pose dataset such as MINI-RGBD [8]. Adding these synthetic occlusions will increase the difficulty of synthetic sequences and bridge the gap between the real and synthetic data used to train the models. Additionally, SMPLify uses a weak-perspective camera in order to estimate both camera and body parameters, in order to do so, it requires the hips and shoulder joints to be present in data. Future work should try to address this limitation for the preterm infant domain for in-the-wild videos given that preterm infants are placed in incubators which can provide significant information regarding camera parameters. Additionally, artefacts such as twisted limbs and mesh interpenetration inherited from the limitations of SMPLify and SMIL are still present.

For occlusion augmentation techniques, future research could address synthetic augmentation techniques for more realistic occlusion sequences as they are highly desirable. We recommend using rendering environments such as Blender in order to create more realistic occlusions. For example, simulating occlusions with cloth physics in order to closely simulate real occlusions as proposed by Achilles et. al. [57] might yield significant improvements in network perfor-

mance. Doing so would provide additional information to the network about how infant movement might look like under severe occlusions, particularly blankets, and increase the quality of the synthetic videos. Future research should additionally look into more sophisticated and complex ways to tackle occlusion modeling in the preterm infant domain. As previously alluded to, Zhou et. al. [59] propose an interesting avenue for occlusion detection as they train a Siamese network to predict heat-map joint locations as well as occlusion maps to erase contaminated features. Future work should investigate whether such a technique would be appropriate for the preterm infant domain. Additionally, this technique could be leveraged in combination with a modified FiDIP framework to encode not only the feature representations of the joints in the same embedding space, but the feature representations of the occlusion maps as well. Studying the performance of such an adversarial framework for occlusion modeling could bear interesting results.

Lastly, future research should address the limitations, shortcomings, and oversights of the current research in regards to TCNs and infant motion. As previously stated in the discussion, re-training the TCNs using a dataset such as MINI-RGBD [8] is likely to yield much more accurate and positive results and their viability for preterm infant pose estimation should be studied. Additionally, future work regarding formulation of the data collection procedure for this domain, as well as additional annotations is required. In order to collect further data on preterm infants, the fine-tuned pose estimator presented in this paper could be leveraged to collect pose data for joints in the head, torso and arms, however it is not advised to use this for points which are often heavily occluded and difficult to predict such as the knees and ankles. As for the TCN component, future improvements could be in regards to training and network fine tuning. The use of infant data has already been previously discussed, however, if there is still a lack of available infant data, we propose to use Human3.6M as indicated in this paper, with some slight alterations. Rather than performing rotation augmentations, a more elegant solution would be to identify the direction which the infant is facing, and rotate the key-points such that the infant appears as if it was standing. Additionally, rather than using the entire image to represent the coordinate space, using a static bounding-box which always has the subject in the center might be more appropriate for the infant domain. Given that we do not want to capture actions, but rather movements, this would allow us to exploit the motion data in an appropriate manner for the preterm infant domain as they do not move around an environment but rather in place.

Lastly, future research should refrain from using a non-static set-up for data collection as it makes the data generation procedure less unconstrained and inaccurate. Preferably the camera should be in the same location, or it should be calibrated prior to each recording to know the extrinsic camera parameters. Knowing these parameters would facilitate the data fitting procedure and would enable us to create more accurate and precise synthetic data. Furthermore, placing the camera above and perpendicularly to the incubators might improve

data quality and ease of annotation; this would reduce the chance of annotation errors as well as facilitate the estimation task and improve the synthetic data fitting procedure results significantly.

Bibliography

[1] H. Blencowe, S. Cousens, M. Oestergaard, D. Chou, A.-B. Moller, R. Narwal, A. Adler, C. Garcia, S. Rohde, L. Say, and J. Lawn, "National, regional, and worldwide estimates of preterm birth rates in the year 2010 with time trends since 1990 for selected countries: A systematic analysis and implications," *Lancet*, vol. 379, pp. 2162–72, 06 2012.

[2] "Preterm birth," 11 2021. [Online]. Available: https://www.cdc.gov/reproductivehealth/maternalinfanthealth/PretermBirth.htm

[3] J. Werth, L. Atallah, P. Andriessen, X. Long, E. Zwartkruis-Pelgrim, and R. M. Aarts, "Unobtrusive sleep state measurements in preterm infants–a review," *Sleep medicine reviews*, vol. 32, pp. 109–122, 2017.

[4] 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*. Springer, 2018, pp. 792–800.

[5] N. Hesse, S. Pujades, M. J. Black, M. Arens, U. G. Hofmann, and A. S. Schroeder, "Learning and tracking the 3d body shape of freely moving infants from rgb-d sequences," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 10, pp. 2540–2551, 2019.

[6] G. Sciortino, G. M. Farinella, S. Battiato, M. Leo, and C. Distante, "On the estimation of children's poses," in *International conference on image analysis and processing*. Springer, 2017, pp. 410–421.

[7] X. Huang, N. Fu, S. Liu, K. Vyas, A. Farnoosh, and S. Ostadabbas, "Invariant representation learning for infant pose estimation with small data," *arXiv preprint arXiv:2010.06100*, 2020.

[8] N. Hesse, C. Bodensteiner, M. Arens, U. G. Hofmann, R. Weinberger, and A. S. Schroeder, "Computer vision for medical infant motion analysis: State of the art and RGB-D data set," in *Computer Vision - ECCV 2018 Workshops*. Springer International Publishing, 2018.

[9] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black, "Keep it smpl: Automatic estimation of 3d human pose and shape from a single image," in *European conference on computer vision*. Springer, 2016, pp. 561–578.

[10] T. Lissauer, A. A. Fanaroff, L. Miall, and J. Fanaroff, *Neonatology at a Glance*, 3rd ed. Wiley-Blackwell, 06 2015.

[11] T. R. Fenton, "A new growth chart for preterm babies: Babson and benda's chart updated with recent data and a new format," *BMC pediatrics*, vol. 3, p. 13, 12 2003.

[12] G. K. Swamy, R. SkjŠrven *et al.*, "Association of preterm birth with long-term survival, reproduction, and next-generation preterm birth," *Jama*, vol. 299, no. 12, pp. 1429–1436, 2008.

[13] J. Volpe, "Brain injury in premature infants: A complex amalgam of destructive and developmental disturbances," *Lancet neurology*, vol. 8, pp. 110–24, 02 2009.

[14] M. Ednick, A. Cohen, G. Mcphail, D. Beebe, N. Simakajornboon, and R. Amin, "A review of the effects of sleep during the first year of life on cognitive, psychomotor, and temperament development," *Sleep*, vol. 32, pp. 1449–58, 11 2009.

[15] L. K. Mestha, S. Kyal, B. Xu, L. E. Lewis, and V. Kumar, "Towards continuous monitoring of pulse rate in neonatal intensive care unit with a webcam," in *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2014, pp. 3817–3820.

[16] W. Verkruysse, L. O. Svaasand, and J. S. Nelson, "Remote plethysmographic imaging using ambient light." *Optics express*, vol. 16, no. 26, pp. 21 434–21 445, 2008.

[17] L. A. Aarts, V. Jeanne, J. P. Cleary, C. Lieber, J. S. Nelson, S. B. Oetomo, and W. Verkruysse, "Non-contact heart rate monitoring utilizing camera photoplethysmography in the neonatal intensive care unit—a pilot study," *Early human development*, vol. 89, no. 12, pp. 943–948, 2013.

[18] N. Koolen, O. Decroupet, A. Dereymaeker, K. Jansen, J. Vervisch, V. Matic, B. Vanrumste, G. Naulaers, S. Van Huffel, and M. De Vos, "Automated respiration detection from neonatal video data." in *ICPRAM (2)*, 2015, pp. 164–169.

[19] Q. Chen, X. Jiang, X. Liu, C. Lu, L. Wang, and W. Chen, "Non-contact heart rate monitoring in neonatal intensive care unit using rgb camera," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2020, pp. 5822–5825.

[20] M. van Gastel, B. Balmaekers, S. B. Oetomo, and W. Verkruysse, "Near-continuous non-contact cardiac pulse monitoring in a neonatal intensive care unit in near darkness," in *Optical diagnostics and sensing XVIII: Toward point-of-care diagnostics*, vol. 10501. International Society for Optics and Photonics, 2018, p. 1050114.

[21] B. Huang, W. Chen, C.-L. Lin, C.-F. Juang, Y. Xing, Y. Wang, and J. Wang, "A neonatal dataset and benchmark for non-contact neonatal heart rate monitoring based on spatio-temporal neural networks," *Engineering Applications of Artificial Intelligence*, vol. 106, p. 104447, 2021.

[22] K. Gibson, A. Al-Naji, J. Fleet, M. Steen, A. Esterman, J. Chahl, J. Huynh, and S. Morris, "Non-contact heart and respiratory rate monitoring of preterm infants based on a computer vision system: A method comparison study," *Pediatric research*, vol. 86, no. 6, pp. 738–741, 2019.

[23] A. Al-Naji and J. Chahl, "Remote respiratory monitoring system based on developing motion magnification technique," *Biomedical Signal Processing and Control*, vol. 29, pp. 1–10, 2016.

[24] S. L. Rossol, J. K. Yang, C. Toney-Noland, J. Bergin, C. Basavaraju, P. Kumar, and H. C. Lee, "Non-contact video-based neonatal respiratory monitoring," *Children*, vol. 7, no. 10, 2020. [Online]. Available: https://www.mdpi.com/2227-9067/7/10/171

[25] A. Heinrich, X. Aubert, and G. de Haan, "Body movement analysis during sleep based on video motion estimation," in *2013 IEEE 15th International Conference on e-Health Networking, Applications and Services (Healthcom 2013)*. IEEE, 2013, pp. 539–543.

[26] X. Long, R. Otte, E. v. d. Sanden, J. Werth, and T. Tan, "Video-based actigraphy for monitoring wake and sleep in healthy infants: A laboratory study," *sensors*, vol. 19, no. 5, p. 1075, 2019.

[27] S. Cabon, F. Porée, A. Simon, B. Met-Montot, P. Pladys, O. Rosec, N. Nardi, and G. Carrault, "Audio-and video-based estimation of the sleep stages of newborns in neonatal intensive care unit," *Biomedical Signal Processing and Control*, vol. 52, pp. 362–370, 2019.

[28] S. Moccia, L. Migliorelli, R. Pietrini, and E. Frontoni, "Preterm infants' limb-pose estimation from depth images using convolutional neural networks," in *2019 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*. IEEE, 2019, pp. 1–7.

[29] C. Zheng, W. Wu, T. Yang, S. Zhu, C. Chen, R. Liu, J. Shen, N. Kehtarnavaz, and M. Shah, "Deep learning-based human pose estimation: A survey," *arXiv preprint arXiv:2012.13392*, 2020.

[30] A. Toshev and C. Szegedy, "Deeppose: Human pose estimation via deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1653–1660.

[31] Y. Chen, Y. Tian, and M. He, "Monocular human pose estimation: A survey of deep learning-based methods," *Computer Vision and Image Understanding*, vol. 192, p. 102897, 2020.

[32] N. Sarafianos, B. Boteanu, B. Ionescu, and I. A. Kakadiaris, "3d human pose estimation: A review of the literature and analysis of covariates," *Computer Vision and Image Understanding*, vol. 152, pp. 1–20, 2016.

[33] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis, "Coarse-to-fine volumetric prediction for single-image 3d human pose," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7025–7034.

[34] X. Sun, J. Shang, S. Liang, and Y. Wei, "Compositional human pose regression," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2602–2611.

[35] 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*, 2017, pp. 2640–2649.

[36] D. Pavllo, C. Feichtenhofer, D. Grangier, and M. Auli, "3d human pose estimation in video with temporal convolutions and semi-supervised training," *CoRR*, vol. abs/1811.11742, 2018. [Online]. Available: http://arxiv.org/abs/1811.11742

[37] 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*, 2017, pp. 398–407.

[38] D. Mehta, S. Sridhar, O. Sotnychenko, H. Rhodin, M. Shafiei, H.-P. Seidel, W. Xu, D. Casas, and C. Theobalt, "Vnect: Real-time 3d human pose estimation with a single rgb camera," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, pp. 1–14, 2017.

[39] 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, pp. 1–16, 2015.

[40] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. V. Gehler, and B. Schiele, "Deepcut: Joint subset partition and labeling for multi person pose estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4929–4937.

[41] G. Varol, J. Romero, X. Martin, N. Mahmood, M. J. Black, I. Laptev, and C. Schmid, "Learning from synthetic humans," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 109–117.

[42] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *European conference on computer vision*. Springer, 2016, pp. 483–499.

[43] Y. Cheng, B. Yang, B. Wang, W. Yan, and R. T. Tan, "Occlusion-aware networks for 3d human pose estimation in video," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 723–732.

[44] G. Ning, J. Pei, and H. Huang, "Lighttrack: A generic framework for online top-down human pose tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 1034–1035.

[45] Q. Dang, J. Yin, B. Wang, and W. Zheng, "Deep learning based 2d human pose estimation: A survey," *Tsinghua Science and Technology*, vol. 24, no. 6, pp. 663–676, 2019.

[46] 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*, 2017, pp. 7291–7299.

[47] A. Nibali, Z. He, S. Morgan, and L. Prendergast, "Numerical coordinate regression with convolutional neural networks," *arXiv preprint arXiv:1801.07372*, 2018.

[48] J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler, "Joint training of a convolutional network and a graphical model for human pose estimation," *Advances in neural information processing systems*, vol. 27, pp. 1799–1807, 2014.

[49] X. Sun, B. Xiao, F. Wei, S. Liang, and Y. Wei, "Integral human pose regression," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 529–545.

[50] D. C. Luvizon, H. Tabia, and D. Picard, "Human pose regression by combining indirect part detection and contextual information," *CoRR*, vol. abs/1710.02322, 2017. [Online]. Available: http://arxiv.org/abs/1710.02322

[51] 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*, 2020, pp. 5386–5395.

[52] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[53] W. Li, Z. Wang, B. Yin, Q. Peng, Y. Du, T. Xiao, G. Yu, H. Lu, Y. Wei, and J. Sun, "Rethinking on multi-stage networks for human pose estimation," *arXiv preprint arXiv:1901.00148*, 2019.

[54] W. Yang, S. Li, W. Ouyang, H. Li, and X. Wang, "Learning feature pyramids for human pose estimation," in *proceedings of the IEEE international conference on computer vision*, 2017, pp. 1281–1290.

[55] H. Zhang, H. Ouyang, S. Liu, X. Qi, X. Shen, R. Yang, and J. Jia, "Human pose estimation with spatial contextual information," *arXiv preprint arXiv:1901.01760*, 2019.

[56] L. Ke, M.-C. Chang, H. Qi, and S. Lyu, "Multi-scale structure-aware network for human pose estimation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[57] F. Achilles, A.-E. Ichim, H. Coskun, F. Tombari, S. Noachtar, and N. Navab, "Patient mocap: Human pose estimation under blanket occlusion for hospital monitoring applications," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2016, pp. 491–499.

[58] J. Wang, E. Xu, K. Xue, and L. Kidzinski, "3d pose detection in videos: Focusing on occlusion," *arXiv preprint arXiv:2006.13517*, 2020.

[59] L. Zhou, Y. Chen, Y. Gao, J. Wang, and H. Lu, "Occlusion-aware siamese network for human pose estimation," in *European Conference on Computer Vision*. Springer, 2020, pp. 396–412.

[60] Y. Cheng, B. Yang, B. Wang, and R. T. Tan, "3d human pose estimation using spatio-temporal networks with explicit occlusion training," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 10 631–10 638.

[61] J. Huang, Z. Zhu, and G. Huang, "Multi-stage hrnet: multiple stage high-resolution network for human pose estimation," *arXiv preprint arXiv:1910.05901*, 2019.

[62] M. Awais, X. Long, B. Yin, C. Chen, S. Akbarzadeh, S. F. Abbasi, M. Irfan, C. Lu, X. Wang, L. Wang *et al.*, "Can pre-trained convolutional neural networks be directly used as a feature extractor for video-based neonatal sleep and wake classification?" *BMC research notes*, vol. 13, no. 1, pp. 1–6, 2020.

[63] I. Sárándi, T. Linder, K. O. Arras, and B. Leibe, "How robust is 3d human pose estimation to occlusion?" *arXiv preprint arXiv:1808.09316*, 2018.

[64] H. Xia and M. Xiao, "3d human pose estimation with generative adversarial networks," *IEEE Access*, vol. 8, pp. 206 198–206 206, 2020.

[65] X. Peng, Z. Tang, F. Yang, R. S. Feris, and D. Metaxas, "Jointly optimize data augmentation and network training: Adversarial data augmentation in human pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2226–2234.

[66] P. Rojtberg, T. Pöllabauer, and A. Kuijper, "Style-transfer gans for bridging the domain gap in synthetic pose estimator training," in *2020 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*. IEEE, 2020, pp. 188–195.

[67] Y. Bin, X. Cao, X. Chen, Y. Ge, Y. Tai, C. Wang, J. Li, F. Huang, C. Gao, and N. Sang, "Adversarial semantic data augmentation for human pose estimation," in *European Conference on Computer Vision*. Springer, 2020, pp. 606–622.

[68] D. Mehta, O. Sotnychenko, F. Mueller, W. Xu, S. Sridhar, G. Pons-Moll, and C. Theobalt, "Single-shot multi-person 3d pose estimation from monocular rgb," in *2018 International Conference on 3D Vision (3DV)*. IEEE, 2018, pp. 120–130.

[69] N. Hesse, G. Stachowiak, T. Breuer, and M. Arens, "Estimating body pose of infants in depth images using random ferns," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2015, pp. 35–43.

[70] N. Hesse, A. S. Schröder, W. Müller-Felber, C. Bodensteiner, M. Arens, and U. G. Hofmann, "Body pose estimation in depth images for infant motion analysis," in *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2017, pp. 1909–1912.

[71] S. Moccia, L. Migliorelli, V. Carnielli, and E. Frontoni, "Preterm infants' pose estimation with spatio-temporal features," *IEEE Transactions on Biomedical Engineering*, vol. 67, no. 8, pp. 2370–2380, 2019.

[72] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. A. Osman, D. Tzionas, and M. J. Black, "Expressive body capture: 3d hands, face, and body from a single image," in *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[73] A. Quattoni and A. Torralba, "Recognizing indoor scenes," in *2009 IEEE conference on computer vision and pattern recognition*. IEEE, 2009, pp. 413–420.

[74] M. Kocabas, S. Karagoz, and E. Akbas, "Multiposenet: Fast multi-person pose estimation using pose residual network," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 417–433.

[75] 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, pp. 1325–1339, 2013.

[76] M. M. Loper and M. J. Black, "Opendr: An approximate differentiable renderer," in *European Conference on Computer Vision*. Springer, 2014, pp. 154–169.

[77] M. Kocabas, N. Athanasiou, and M. J. Black, "Vibe: Video inference for human body pose and shape estimation," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Appendix A

Synthetic Data

A better perspective regarding the variation of the augmented data can be seen in Figure A.1. This figure shows 8 frames from different synthetic sequences in the SPIS dataset, each frame has a different background, different viewing angle and different texture. These images are not exhaustive of the variability of the dataset, but provide good insight in regards to how different the infants and images are in the dataset. Due to the large selection of backgrounds, infant textures and viewing angles the variability of the data is relatively high. Different fitted sequences from the same real infant sequence always look different and provide a challenging amount of variation. This visual variability is key in training the pose estimator, in particular the different infant textures are particularly useful for training an in-the-wild model. Given the visual diversity textures, we can ensure that the trained pose estimator is capable of predicting poses for groups of different ethnic groups.

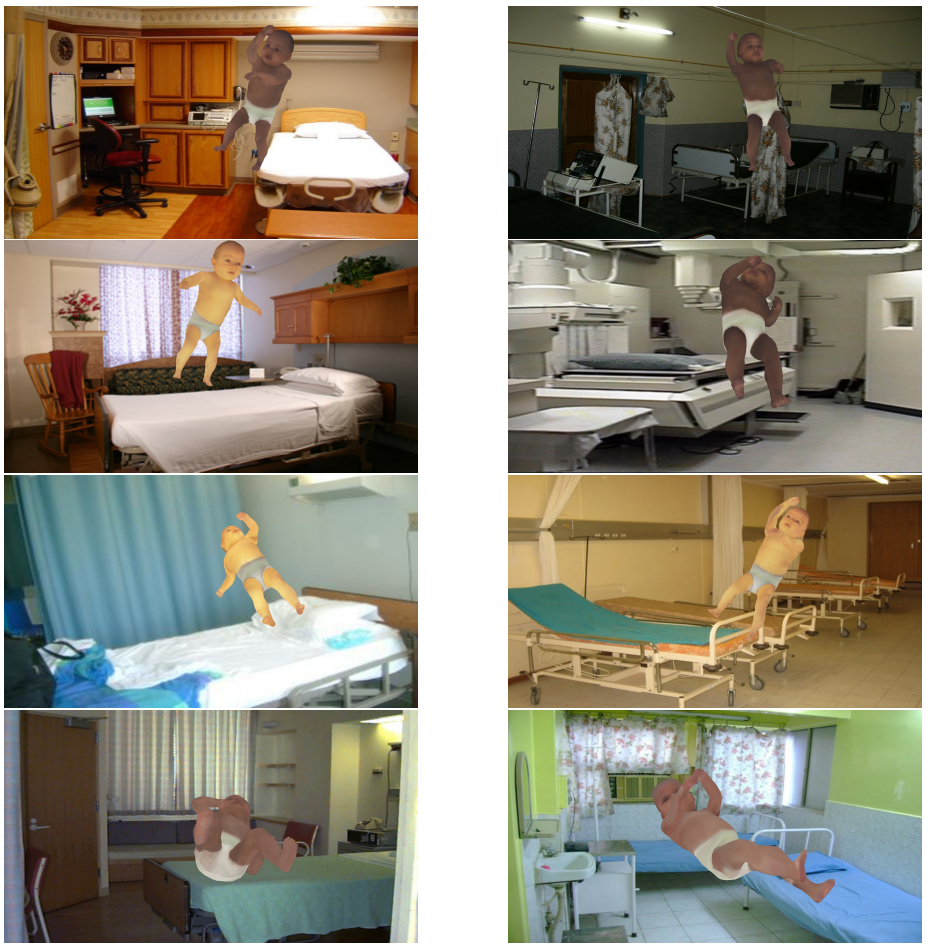


Figure A.1: Synthetic image samples from the SPIS dataset. Each image has a different background, camera perspective and infant texture.

Appendix B

TCNs

B.1 Pre-training Results

Tables B.1 and B.2 show the results of the TCNs on the complete SPIS dataset without rotating augmentations during training. The results clearly indicate a significant drop in performance, and the network outputs are extremely noisy. These results also demonstrate why using PCK as a sole metric can be extremely misleading. These results indicate why rotation augmentations were necessary, as show by the increased performances presented in Section 5.3.3

Metric \ Condition	TCN-B	TCN-OA	TCN-DA	TCN-DA-FT	TCN-DA-OA	TCN-DA-OA-FT
Nose MSE	163.4	188.7	122.4	149.4	135.8	178.8
Eyes MSE	143.0	176.7	118.1	168.2	130.2	139.6
Ears MSE	182.5	199.2	116.6	133.1	155.4	115.7
Shoulders MSE	158.0	200.4	87.6	124.4	142.2	95.3
Elbow MSE	203.2	280.7	160.1	180.1	251.1	181.2
Wrist MSE	230.4	298.2	192.7	231.9	293.8	216.6
Hip MSE	280.1	354.8	157.0	162.9	306.0	142.9
Knee MSE	232.4	371.2	261.2	293.9	352.7	255.2
Ankle MSE	265.5	435.6	274.2	327.5	429.3	307.4
Total MSE	209.03	283.66	168.1	199.6	250.4	178.8

Table B.1: Total MSE and per joint MSE in pixels for all ablation conditions.

Metric \ Condition	TCN-B	TCN-OA	TCN-DA	TCN-DA-FT	TCN-DA-OA	TCN-DA-OA-FT
Nose PCK	45.17	34.1	43.95	48.40	39.04	43.20
Eyes PCK	53.72	31.12	54.28	58.70	36.31	53.20
Ears PCK	35.60	20.40	45.52	52.10	21.65	54.40
Shoulders PCK	41.36	17.56	53.50	58.37	17.50	63.69
Elbow PCK	17.55	5.51	17.41	29.01	4.31	29.29
Wrist PCK	9.54	5.70	11.31	29.86	4.85	35.50
Hip PCK	19.86	2.48	24.98	43.24	3.26	52.13
Knee PCK	24.80	14.50	30.58	43.02	20.37	46.79
Ankle PCK	11.09	3.80	16.19	38.83	4.50	41.10
Total PCK	27.78	13.40	32.44	45.79	15.56	46.80

Table B.2: Total PCK@0.2 and per joint PCK@0.2 for all ablation conditions. Note that it is abbreviated to PCK on the table.

B.2 Upper-body results

Metric \ Condition	TCN-B	TCN-OA	TCN-DA	TCN-DA-FT	TCN-DA-OA	TCN-DA-OA-FT
Nose MSE	89.5	70.2	62.8	51.3	49.4	49.1
Eyes MSE	86.3	69.1	60.7	48.8	52.9	47.6
Ears MSE	94.6	81.3	62.4	52.8	59.1	51.0
Shoulders MSE	89.6	77.9	44.8	37.3	40.5	37.1
Elbow MSE	109.0	90.6	57.6	50.6	48.0	46.6
Wrist MSE	123.9	110.7	101.9	81.8	80.5	71.1
Total MSE	98.8	83.3	65.0	53.8	55.1	50.4

Table B.3: Total MSE and per-joint MSE in pixels for the upper-body joints on the complete SPIS test split for all ablation conditions.

Metric \ Condition	TCN-B	TCN-OA	TCN-DA	TCN-DA-FT	TCN-DA-OA	TCN-DA-OA-FT
Nose PCK	62.8	64.5	62.1	68.1	65.6	60.6
Eyes PCK	63.8	63.6	60.9	70.0	63.4	65.4
Ears PCK	41.4	35.9	42.2	53.6	36.4	52.4
Shoulders PCK	59.6	56.0	64.6	67.2	62.4	69.5
Elbow PCK	51.1	50.8	59.5	60.1	58.3	61.2
Wrist PCK	39.7	37.8	46.1	46.2	45.3	45.8
Total PCK	53.1	51.4	55.9	60.9	55.2	59.2

Table B.4: Total PCK@0.2 and per-joint PCK@0.2 for the upper-body joints on the complete SPIS test split. for all ablation conditions. Note that it is abbreviated to PCK on the table.

Metric \ Condition	TCN-B	TCN-OA	TCN-DA	TCN-DA-FT	TCN-DA-OA	TCN-DA-OA-FT
Nose MSE	168.5	132.5	138.9	111.3	102.6	102.7
Eyes MSE	161.5	130.3	133.4	104.9	111.8	99.2
Ears MSE	172.9	147.8	119.4	102.8	110.6	96.5
Shoulders MSE	180.2	155.1	88.1	71.0	74.1	69.7
Elbow MSE	231.2	192.0	119.6	105.0	94.7	94.0
Wrist MSE	245.9	222.4	219.4	176.2	165.1	146.5
Total MSE	193.3	163.4	136.5	111.9	109.8	101.4

Table B.5: Total MSE and per-joint MSE in pixels for the upper-body joints on the real data SPIS test split for all ablation conditions.

Metric \ Condition	TCN-B	TCN-OA	TCN-DA	TCN-DA-FT	TCN-DA-OA	TCN-DA-OA-FT
Nose PCK	36.2	42.6	26.5	38.2	39.2	24.8
Eyes PCK	35.9	40.5	24.6	40.2	31.4	31.8
Ears PCK	24.5	21.5	29.6	30.6	20.1	31.7
Shoulders PCK	38.9	42.8	41.3	49.1	48.9	53.0
Elbow PCK	22.9	24.8	31.9	28.8	31.6	33.0
Wrist PCK	12.7	9.9	14.1	11.6	15.1	15.3
Total PCK	28.5	30.4	28.0	34.1	31.1	31.6

Table B.6: Total PCK@0.2 and per-joint PCK@0.2 for the upper-body joints on the real data SPIS test split. for all ablation conditions. Note that it is abbreviated to PCK on the table.

These results indicate the total MSE and total PCK of the models for only the upper-body joints of the body; the hip, knee, and ankle joints are not considered. The results presented here remain consistent with the results presented in Section 6.1.2; albeit these results paint show that if we only consider the upper body joints, the overall performance of the models is slightly more positive. Table B.3 indicates that TCN-DA-OA-FT still had the lowest total MSE for the SPIS test set split, and Table B.4 indicates TCN-DA-FT had the highest PCK, as seen in the non-filtered joint data. Tables B.5 and B.6 contain the MSE and PCK scores for the models of the upper joints for the real data split only. These tables provide the same results as the non-filtered tables shown in Section 6.1.2, where TCN-DA-OA-FT still had the lowest total MSE while TCN-DA-FT had the highest total PCK score. By comparing the results between the full joints and the upper-body joints, we can reach the conclusion that the performance does not improve enough to consider using the trained models only for the upper body joints. The results indicate that the architecture does not produce the quality of the results required for infant motion analysis.