REALISTIC OCCLUSION AUGMENTATION FOR HUMAN POSE ESTIMATION

Amin Ansarian

A thesis

in

The Department

of

Electrical and Computer Engineering

Presented in Partial Fulfillment of the Requirements

For the Degree of Master of Applied Science in Electrical and Computer Engineering

Concordia University

Montréal, Québec, Canada

July 2021

Concordia University

School of Graduate Studies

This is to certify that the thesis prepared

By: **Amin Ansarian**

Entitled: **Realistic Occlusion Augmentation for Human Pose Estimation**

and submitted in partial fulfillment of the requirements for the degree of

Master of Applied Science in Electrical and Computer Engineering

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the final examining commitee:

Dr. Krzysztof Skonieczny _____ Chair

Dr. Charalambos (Charis) Poullis _____ External Examiner

Dr. Krzysztof Skonieczny _____ Examiner

Dr. Maria A. Amer _____ Supervisor

Approved by _____

Dr. Akshay Kumar Rathore, Graduate Program Director

_____ 2021 _____

Dr. Mourad Debbabi , Dean

Faculty of Engineering and Computer Science

Abstract

Realistic Occlusion Augmentation for Human Pose Estimation

Amin Ansarian

Occlusion occurs naturally in a high percentage of real-world images. Handling occlusion has been a difficult challenge in human pose estimation methods, specially those using CNN. A main reason is the a lack of a proper dataset with an actual focus on realistic occlusion, prompting researchers to create datasets with synthetic (bounding-box based) occlusion, as a means of data augmentation. In this thesis, we investigate how to increase learning through data preparation (i.e., data-centric approach). To this end, we introduce a new realistic data augmentation approach built on top of an original (base) dataset (e.g., Human3.6m and MPI-INF-3DHP) that tackles this issue, creating realistic samples similar to those found in the wild. Arguing that CNN models pay higher attention to local as opposed to global features, we define occlusion levels, process a large set of occluder objects from different categories, augment them adaptive to the joints types and to the size of the human subject, and effectively blend those occluders within the original RGB image from the base dataset. We, then, test top-performing CNN-based 2D and 3D human pose estimation models with and without our occlusion-augmented datasets (*RealPose*). Our experiments show that a significant drop in accuracy of these CNN models under occlusion. When we then train them on RealPose, we observe a major increase in accuracy under occlusion, without any change to the models themselves. Achieved results indicate the effectiveness of the proposed data augmentation method in tackling the occlusion issue both in the 2D and 3D models, with a significantly much more

accuracy increase of the 3D models. We have trained and tested the models under different dataset combinations such as "training on the original dataset but testing under the augmented dataset" or "training and testing with the mixed original and augmented dataset". A significant outcome is that, our data-centric approach results in a higher accuracy of CNN models trained under occluded samples but tested under the original (not-occluded) samples, indicating the model achieves a higher understanding of the dependency of different joints induced to it. Proposed approach is dataset and network independent, i.e., researchers can apply our approach (using its open-source code) to any dataset and feed the result to any human pose estimator.

Acknowledgments

First and foremost, I am extremely grateful to my supervisor, Dr. Maria A. Amer, for her invaluable advice, continuous support, and unwavering patience during my study. Her immense knowledge and plentiful experience have encouraged me in all the time of my academic research and daily life. This work would not have come into the shape it is today without her support.

I also wish to extend my special thanks to my wife, Attefe, for her unending and unconditional support throughout the entire course of my studies for I would have never succeeded without her invaluable presence. I dedicate this entire thesis to her.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Human Pose estimation has many applications in human action recognition and Human-Computer Interaction. The purpose is to estimate a certain number of joints in the two-dimensional space from an RGB image, which due to factors such as the high variation among human anatomies, different and unexpected occlusion, has deemed to be a sophisticated task. With the emergence of complex deep-learning-based models, especially those using Convolutional Neural Networks (CNN), a high level of advancement has been obtained [8, 56, 57, 60, 49], resulting in models deployed in industrial applications.

A major challenge for human pose estimation is occlusion [38, 39, 40, 52, 12, 49, 48], meaning one or more joints are hidden from the camera due to various factors such as self-occlusion, camera zoom, angle or obstruction by random objects. This issue causes divergence in both training and testing, often resulting in poor accuracy by the CNN model. To handle occlusion, generally, researchers have followed two approaches: 1) directly target occlusion of the human subject by designing a customized loss function or taking anatomy information into account[8, 9, 53, 6] and 2) incorporating

data augmentation during the training of the model [61, 51, 10, 64, 11, 45, 9, 55, 20]. The problem with these approaches is a lack of realism that usually occur in the wild.

1.2 Problem Statement

There is an important discussion in the research community [42, 41] about data-centric versus model-centric learning. In this thesis, we show that learning can be increased through well-shaped data preparation. To this end, we propose a data augmentation approach that results in a realistic occlusion dataset that incorporates real-world objects for occluding any specific joint of a human subject (sample) and provides the occlusion label for that specific joint. By doing so, we move further closer to the actual occlusion in real life, providing an opportunity for researchers to train occlusion-aware models on the said dataset. The significant need for a well-shaped occlusion augmented dataset is due to the fact that no dataset with such level of reality and detail exists in literature that would allow researchers to test their methods under realistic occlusion. Moreover, since the base datasets we use are the Human3.6m and the MPI-INF-3DHP, which both provide the full skeleton annotations for each human subject in a video frame, the proposed (augmented) dataset will be fully compatible with the original base dataset, not affecting the model (without occlusion) training. The contributions of this thesis are: i) a data augmentation approach that works on any available human pose estimation dataset, providing close to realistic occlusion; ii) extensive testing of the functionality of the approach on both 2D and 3D human pose estimation CNN models in the presence of the occlusion problem; iii) significant improvement of the accuracy of 2D and 3D human pose estimation models under occluded and not occluded subjects, itself a breakthrough to be considered when introducing data augmentation for training.

1.3 Summary of Contributions

The main objective of this thesis is handling the occlusion problem in Human Pose Estimation. To this end, we propose a novel data augmentation method to create an occlusion dataset (called RealPose) that incorporates contextual information close to the cases that might occur in real-world cases. In order to verify the effectiveness of our dataset, we implement two 2D human pose estimators and evaluate them under different conditions, namely, original not-occluded datasets, occluded datasets and a novel mixed dataset. We further extend our work to 3D human pose estimation to assess the functionality of RealPose and implement two 3D human pose estimators and then, evaluate their performance under the same conditions. Furthermore, we confirm the validity of our occlusion augmentation method by extending our work to a second dataset. As a result of our work, the paper [1] "Realistic Augmentation for Effective 2D Human Pose Estimation under Occlusion" has been accepted for publication at the IEEE ICIP on May 20, 2021.

1.4 Thesis Outline

In Chapter 2, a comprehensive analysis of related works is presented. We analyze and discuss multiple works in Human Pose Estimation, Works that deal with occlusion without augmentation methods and finally, works that handle occlusion by augmentation methods.

In Chapter 3, we present our occlusion augmentation approach. First we introduce the datasets that are used in thesis and discuss their specifications and the reasons they were selected. Then, we elaborate on our approach and its details. At the end, we discuss the various hyper-parameters used in our method.

In Chapter 4, we begin by discussing the implementation details of the candidate models. Then, we present the 2D pose estimation results with and without our

occlusion augmentation method. Next, we extend the work to 3D pose estimation and discuss the results. We finalize the chapter by discussing several points regarding the evaluation.

In Chapter 5, we summarize our contributions and present future work to conclude the thesis.

Chapter 2

Related Works

In this section, we first provide a review of state-of-the-art to human pose estimation and then we discuss two categories of occlusion-aware pose estimation methods: those without and those with data augmentation.

2.1 Human Pose Estimation - a review

Human pose estimation can be classified into monocular-image based estimators [49, 21, 8], multi-view based estimators [27, 23, 44, 25], single-person pose estimator [6, 43, 11], and multiple-persons pose estimators [8, 21] (Recent survey papers are [7, 18, 33, 54]). Furthermore, estimated poses can be 2D or 3D, where 3D poses can be extracted directly from 2D images [51] or based on estimated 2D poses [43, 6, 11]. Multi-view methods receive images of the human subject from different angles and using various methods [27, 23, 44, 25], fuse them together, producing a more accurate pose compared to that of monocular images [28]. However, these methods are expensive, both in terms of computational cost and their hardware requirements to capture the multi-view images. Also, fusing multiple images together is an extra level of complexity added to the algorithm, making it less efficient compared to one based on commonly available monocular images.

Deep learning-based 2D human pose estimators have shown significant results recently [39, 40, 4, 5, 8, 21, 49, 37]. However, 3D pose estimation is still a challenging task due to the ambiguous third dimension and occlusion. The work of [43] studies the application Dilated Convolutions and Temporal neural networks in the task of 3D pose estimation, achieving a favourable result in addition to a relatively low number of parameters.

Looking at the human skeleton as a graph-like data, [3] and [62] aim to utilize the connectivity of the nodes and local-to-global features by creating Convolutional Graph Neural Networks to extract features on different levels of representation. While both works effectively manage the regression of the 3D joints by reforming the data into a graph, they do not fully exploit the functionality of these models by introducing occlusion into the data.

The work of [11, 10], directly address the case of occlusion in 3D pose estimation. [11] creates an occlusion-aware network by manipulating the extracted 2D heatmaps of the joints and enforcing multiple constraints and penalties on the loss functions. [10] completes the work of [11] by creating multi-scale features for pose estimation to capture fast motions in videos resulting in failures.

Our augmentation approach has been tested for 2D single-pose estimation directly from a 2D RGB image [8, 21] and for 3D single-pose estimation based on these extracted 2D poses [43, 6]. However, it can be extended to 3D pose estimation directly from a 2D image or to multiple-pose estimation by providing a bounding box for each human subject in the RGB image and occluding each separately.

2.2 Occlusion-aware approaches without augmentation

Works using this approach handle occlusion without augmenting the training set with additional data and instead integrate the information regarding the structural information of the human body to the training. The work of [8] adds a sub-network called *RefineNet* to their architecture that collects features extracted from multiple resolutions that accounts only for *hard-examples* which are ignored by the main network. The works of [9, 53] use a spatio-temporal discriminator based on body structure to assess the validity of the predicted pose while the work of [6] explores the role of 2D visibility scores, indicating the probability of the presence of the joint, as a complementary information to handle occlusion.

2.3 Occlusion-aware approaches with augmentation

Works of this category handle occlusion by adding more samples to the training set that contain occlusion-related information [61, 51, 10, 64, 11, 30, 2, 58, 14]. The method in [19] aims to jointly optimize data augmentation and network training where the augmentation is done by deforming a human template model with the ground truth in order to improve the detection accuracy for occluded cases. The authors of [17] augment the human subject with additional segmented body parts in an adversarial way and use a discriminator to learn the correct pose. However, the added cost in model size and training time makes the method not attractive for data augmentation. In [34], the authors propose data augmentation by applying complex image transformations in image classification; the method is computationally very complex and not scalable for large datasets. The work of [30] generates synthetic 3D poses based on an initial training set and project them back to 2D using camera

parameters, creating additional 2D-3D pairs for generalizing the model. The work of [58], similarly, generates 2D synthetic 2D poses by projecting a 3D pose on virtual camera planes and extracts the 2D poses by conditional sampling. The work of [11] also creates additional 3D samples by creating a synthetic "Cylinder-Man" and rotating the camera to create several additional samples for training.

The most common data augmentation approach in the literature, however, is random erasing[64] or cutout [14], where a random part of the image is masked by a uniform patch. This approach is unrealistic in the sense that occlusion and obstruction do not happen in real-world samples in this way. Furthermore, based on an interesting observation by [16, 15] it is argued that CNN-based object detectors (and we argue human pose estimator as well), pay a higher amount of attention to local as opposed to global content. For instance, when detecting an exemplary joint, the joint's visual features and its integration into the background image are much more important than the joint's global location. As networks get more complicated, this" context-aware" characteristic of CNN networks gets more sophisticated. Therefore, the model will learn the "black-patch" instead of the missing joint, failing during inference on real-world cases.

The authors of [45] investigated the accuracy of 3D humane pose estimation methods under occlusion, which was simulated through random box erase and random addition of objects anywhere to the input RGB image; the objects can be from any, even unrealistic, category (such as airplane) and they may appear at random locations in the image, not necessary covering portion of the human subject.

Since pose estimation models typically estimate poses (2D and 3D) based on the bounding box of the subject, occlusion needs to be defined based on bounding box. Another important difference to [45] and all other related work, is that we show the effectiveness of our approach and the pose estimation models under several combination of samples: occlusion, no-occlusion, mixed occlusion/no-occlusion, etc.,

as detailed in Table 2.

In non-pose estimation literature, the work of [29] studies the vulnerability of Deep CNNs to occlusion in object detection and proposes mixing DCNNs with compositional models along with synthetic data generation with samples similar to those of ours to handle object detection under occlusion. Similarly, the authors of [63] propose to exploit contextual information using LSTM [24] networks as well as occluded samples, created using the random erasing technique, to outperform conventional CNN networks for the task of object classification.

Chapter 3

Proposed Occlusion Augmentation

Occlusion occurs when the target object (here human subject) is occluded by another visual object (occluder) that can be of any category (such as Cat, Laptop, or Pants). Inspired by the observation in [16, 15] that CNN models pay higher attention to local context as opposed to global features, we define the concept of "occlusion level," to indicate how much of the human subject body, specifically, how many joints of the total joints J are occluded. We propose 3 local levels of occlusion: heavy, medium, and light, where respectively J_l joints of the total joints J are occluded. We use one occluder object per joint for occlusion (see examples in Figure 2(f)-(h)).

3.1 Base Datasets for Augmentation

The majority of 2D human pose estimators are trained on the COCO dataset [32] due to the high number of samples, high variation in the images and partial occlusion present in the dataset. While this will result in accurate classification [60, 8, 49, 31] of the joints that are present, it makes the skeleton reconstruction impossible, which is crucial for tasks such as human action recognition (a major application requiring human pose estimation) since the skeleton is incomplete and only includes a portion of joints that are visible in the image. Due to this issue we decided to choose, as a

base dataset, one with complete annotations for all samples.

Several datasets are used for 3D human pose estimation [26, 36, 47, 50]. We selected the popular datasets MPI-INF-3DHP [36] and Human3.6m[26] that provide complete skeleton for every image, which is a raw foundation to test the usefulness of our approach. In these datasets, each subject performs different scenarios and each scenario consists of several thousand of samples.

The Human3.6m dataset [26] consists of 3.6 million samples collected with the help of 11 real-world actors across 15 scenarios. This is the largest available dataset that is mostly used in 3D human pose estimation tasks, but it provides 2D annotations as well. We use subjects 1,5,6,7,8,9,11, per standard practice in literature, that include roughly two million samples recorded at 50 frames/second (fps) from 8 different angles. To reduce redundancy in the dataset, we downsample each video to 10fps by sampling 1 in 5 consecutive frames, resulting in precisely 422,420 frames used for creating the base dataset.

The MPI-INF-3DHP [36] dataset consists of 8 subjects performing various actions in two different sequence, each using 14 cameras from different angles. The subjects wear different sets of clothing in each sequence to increase generalizations. The test set for this dataset takes place both indoors and outdoors to test the robustness and generalization of the model. In order to reduce redundancy in the dataset, we downsample the frames by a factor of 5.

Since the Human3.6m and MPI-INF-3DHP datasets provide different categories, number and order of joints, we unify them using a fixed set of joints $J = 17$ with the same skeleton arrangement, as displayed in Figure 1, to facilitate occlusion augmentation and model training.

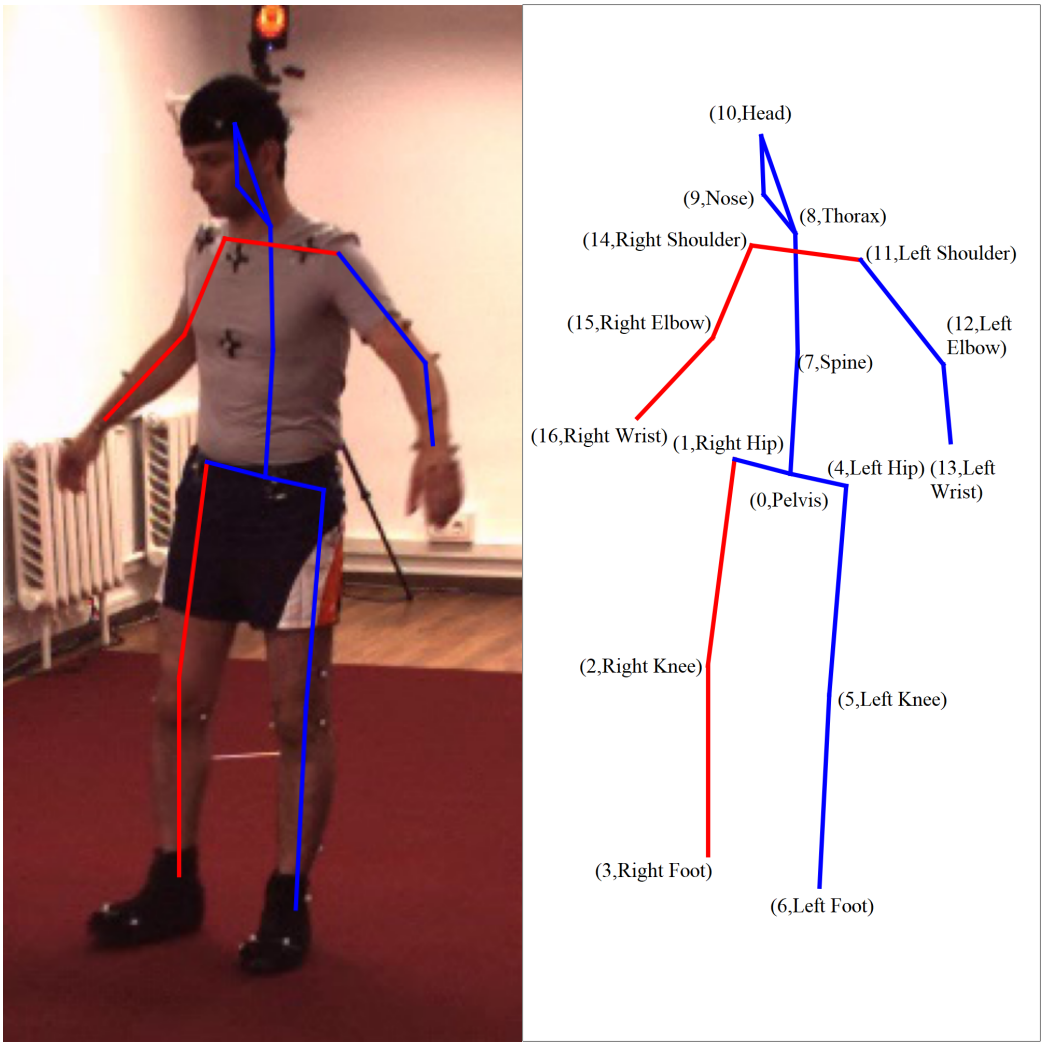

Figure 1: Illustration of our unified pose arrangement applied to the base datasets Human3.6M and MPI-INF-3DHP.

3.2 Proposed Approach

Our approach consists of three main steps: i) process the occluder objects; ii) perform the occlusion scenario; iii) blend the occluders with the human subject in the RGB image.

For the first step, we obtained the masks from the COCO 2017[32] training dataset due to its wide variety of categories, however, we remove categories such as boat, horse or vehicle, etc. from the list of eligible categories due to either not compatible size

or not realistic match with the subjects in the base datasets (here, Human3.6m and MPI-INF-3DHP). A researcher can, however, easily add such categories if needed for the application. We use the features of each occluder object (its RGB image, bounding box, binary mask, area, and category) to facilitate creating the occlusion scenario (i.e., blending into the input RGB image).

In the second step, we perform occlusion on a batch of n images from the original RGB samples. For each of the 3 occlusion levels (light, medium, and high) we assign J_l joints to occlude, i.e., in each image J_l joints should be occluded with J_l occluders. The occlusion levels are balanced using weights w_l, meaning, for a batch of n images, $w_1\%$ are lightly occluded, $w_2\%$ are moderately occluded, and $w_3\%$ are heavily occluded. In order to increase the realistic nature of our augmented dataset (RealPose), we limit certain occluder categories to certain joints category of the body. Hence, each occluder category will belong to a limited number of joints category. As an example, a potted plant will not appear on the upper body and is limited to the bottom joints. Therefore, for a certain occluder object with a category we have a number of potential candidate joints that we can assign it to; to do this we randomly select a joint j from the specific joints to which the object category is limited. We do not select the joints 8,9,10,11 and 14, i.e., the shoulders, neck and head, due to the less probable occlusion in real-world cases. Thus, for occlusion, in total we consider in total 12 joints out of the 17 available. Note that for training we still use 17 joints; the 12 joints are only used for occlusion. We define 4 joint categories and occluders as follows:

- **Only Foot**: Cat, Dog, Fire hydrant, Bench, Chair, TV, Potted plant, Laptop
- **Only Hand**: Banana, Apple, Sandwich, Orange, Broccoli, Carrot, Hot dog, Pizza, Donut
- **Upper Body**: Bird, Sports ball, Bottle, Wine glass, Cup, Bowl, Cake, Mouse,

Remote, Keyboard, Cell phone, Scissors, Teddy bear, Hair drier, Toothbrush

- **All Body**: Backpack, Umbrella, Handbag, Suitcase, Frisbee, Book, Clock, Vase, Laptop

There are overlaps in these categories: All Body includes all of the joints and Upper Body includes Hands as well. This overlap is due to the specific nature of some categories, for example, Books can be on any part of the body while a Fire hydrant will never be on the hands. We randomly select occluders from 64,234 objects across 40 different object categories from the COCO dataset. Given the large number of input RGB samples (422K in the Human3.6m dataset), the majority of the occluder objects are selected by our method. Each Joint category consists of several joints which are illustrated in Table 1:

Table 1: Annotation of Joints and Joint categories

Joint Category	Joints
All Body	All Candidate Joints
Upper Body	Pelvis, Right Hip, Left Hip, Spine, Left Elbow, Left Wrist, Right Elbow, Right Wrist
Only Hand	Left Elbow, Left Wrist, Right Elbow, Right Wrist
Only Foot	Right Knee, Right Foot, Left Knee, Left Foot

Since the area A_h of the human bounding box can be much larger or much smaller than the area A_o of the occluder object, we need to resize the occluder. In order to increase variety in the occlusion dataset (and hence reduce bias), we randomly select the ratio of the occluder object area to the human bounding box area $X = A_o/A_h$ from certain ranges $[a, b]$. The upper b and lower a bounds of these ranges are chosen based on the joint categories, for example, occluders that are assigned to the feet and legs will have a larger size than those assigned to only hands. We resize each occluder using the scale factor SF

$$SF = \frac{A_h \times (X \sim U(a, b))}{A_o}, \tag{1}$$

where, A_h is the area of the human bounding box, A_o is the area of the occluder object, and X randomly selected from the uniform distribution denoted by U, with lower bound a and upper bound b.

SF is a positive real number. And hence, if the width of the occluder image is W and its height H, then the width of the resized occluder becomes $SF \times W$ and its height $SF \times H$, each rounded to the nearest integer, e.g., if $SF = 3.865$, and the occluder size is 15×15, then the output size would be 58×58, while if $SF = 3.8$, the output size would be 57×57. As a consequence, if $SF > 1$ the occluder image will be interpolated; otherwise it will be downsampled (to resize we use the OpenCV function cv2.resize). To our best knowledge no data augmentation method proposes a procedure to compute an adaptive scale factor. Related works such as [46, 59, 45] use a fixed scale factor. Our adaptive scale factor increases the realistic nature of our occlusion augmentation; the randomness we use in computing the scale factor increases variety in our dataset and hence helps reducing overfitting.

At the third and final step of our method, in order to smoothly blend each occluder (from the COCO dataset) on an input RGB image of the base dataset (here Human3.6m and MPI-INF-3DHP) we propose the following. Let I be the input image, O the image of the occluder, M the binary mask of O with pixels either 0 or 255, and B a binary structuring element. To blend O into I, we first apply morphological erosion to M as in $M_b = M \ominus B$, where B is a 8×8 disk. Then, we get $E = M - M_b$, resulting in an edge mask E that is 255 on the edges and 0 otherwise. Next, we mark the respective edge pixels of E in the mask M with special label (e.g., 191). With this, we get a gray-level mask M_o: 0 (not-object), 191 (edge), and 255 (object); we divide M_o by 255 to get a non-integer mask of 0, 0.75, 1. Finally, we blend O into I as in

$$I_o = (1 - M_o) \cdot I + M_o \cdot O \tag{2}$$

We apply blending to all of the three channels of the input RGB image. With equation 2, at edges of the occluder, we get a weighted summation of both the input RGB image and the RGB image of the occluder; inside the boundaries of the occluder we only add the occluder RGB values; and finally outside the occluder, the input RGB image remains unchanged.

3.3 Parameter Analysis

The parameters of our augmentation approach are J_l the number of joints per each of the three occlusion levels (light, medium, and heavy), w_l the weight of each occlusion level across the whole database, and a, b the area ratio bounds (see equation 1).

The number of joints J_l to occlude depends on the resizing coefficient range $[a, b]$, meaning, if the range is too wide (e.g., $a = 30, b = 50$), setting J_l too high (such as 5 or 6 out of 12) will cause out-of-proportion occlusion, covering the entire human subject, while setting J_l too small along with a narrow range (e.g., $a = 5, b = 10$) will result in insignificant occlusion, having no effect on the training. Based on empirical experiments and 12 selected joints, we set $J_1 = 2, J_2 = 4$, and $J_3 = 6$ for light, medium and heavy occlusion, respectively. Using a fixed scale factor SF for resizing, we observed that lower or higher J_l will result in either insignificant occlusion or completely blocking the sample and thus, poor training.

For the area ratio range $[a, b]$, first experiments have shown that any value over 30% of the human subject bounding box area will result in too much occlusion, creating a high amount of ambiguity for the potential pose. Second, since in real-world scenarios lower human body parts (such as foot) is more occluded than upper part (such as head), the range should depend on the joint category. In consequence, we distribute the ranges $[a, b]$ per joint category as follows:

- Only Foot: 0.15 to 0.30

- Only Hand: 0.075 to 0.15
- Upper Body: 0.09 to 0.18
- All Body: 0.075 to 0.15

Thus, each range indicates a uniform distribution from which a random value is selected which indicates the final ratio (see equation 1) of the occluder object area to the human subject area.

As for the weight w_l of occlusion levels on the dataset, the key idea is to balance the number of heavily occluded samples, which are the most difficult, and the number of lighter occluded samples which will preserve the ability of the pose estimation model to detect the skeleton and maintain training, avoiding divergence during the training phase. As a result, we set $w_1 = 15\%$, $w_2 = 50\%$, and $w_3 = 35\%$ for light, medium, and heavy occlusion, respectively. We optimized the values of the discussed parameters experimentally, however, they can be adapted to specific needs of an application by a researcher using our code. Figure 2 illustrates output samples using our approach.

The proposed joint-occluder category constraints maybe viewed as a high degree of manual intervention to augment occlusion. We argue that such constrained augmentation is important for the following three reasons. First, our objective is to make the occlusion augmentation as realistic as possible (for example, not including objects such as bus or airplane in an indoor environment). Second, contextual information is of high importance for learning [16], and thus, we argue that adding joint-occluder context (meaning what occluder category to what joint to occlude) will improve learning of the CNN model trained with our occlusion dataset. Third, there is an important discussion [41, 42] in the AI research community about data-centric versus model-centric advantages [42, 41]. We argue that our data augmentation, resulting in the RealPose dataset contributes to data-centric approaches. Alternatively, we could have created a dataset with minimal intervention, for example, by occluding

any joint of any occluder object. The effect of this, however, would be a reduced amount of contextual information and reduced realism. To show this, we implemented a method where we augment any joint with any occluder object and results show that lower accuracy of pose estimation is achieved compared to our constrained augmentation method.

Figure 2: Output of our data augmentation approach applied to different subjects and scenarios of the Human3.6m dataset (Top and middle rows) and MPI-INF-3DHP dataset (Bottom row). Images (f), (g) and (h) are samples of light, medium and heavy occlusion, respectively.

Chapter 4

Experimental Results

To test the effect of our data augmentation accurately, first, we train a pose estimation model on the original (base) datasets without occlusion augmentation to have a base error for comparison. Then, we test the model separately on both original (without occlusion) and the augmented dataset (RealPose) to see if the expected higher estimation error occurs due to occlusion. Finally, we train and then test the model using the augmented dataset to observe if the error decreases. We do this for different dataset combinations OrgOrg, OrgOcc, OccOrg, MixOrg, OccOcc, MixMix, as detailed in Table 3.

It is important to note that pose estimation models, for example [8, 21, 49, 35], consist of two main parts, a large size feature extractor such as ResNet [22] (often referred to as "backbone") and the estimation network (often referred to as "head" or "main network"), which transforms features received from the backbone to features of the pose estimation (joints). The feature extractor of these models are all trained on large size datasets such as ImageNet [13], which consists of one million training image samples and 1000 object categories. It is common practice that during CNN model training, the feature extractor is not re-trained and only the head of the network is trained.

4.1 Implementation Details

Our primary evaluation metric is the Mean Per Joint Positional Error (MPJPE) in millimeter, which is the per joint Euclidean distance. MPJPE is calculated as follows [26]

$$MPJPE = \frac{1}{T}\frac{1}{J-1}\sum_{t=1}^{T}\sum_{j=1}^{J-1}||(K_j^{(t)} - K_{root}^{(t)}) - (\hat{K}_j^{(t)} - \hat{K}_{root}^{(t)})||_2. \quad (3)$$

In this equation, the joint values are normalized based on a reference joint, which is set to pelvis as per standard practice, hence, the number of joints is $J-1$, meaning 16. Furthermore, T is the total number of samples, $K_j^{(t)}$ is the ground truth joint, $K_{root}^{(t)}$ is the ground truth pelvis, $\hat{K}_j^{(t)}$ is the predicted joint, and $\hat{K}_{root}^{(t)}$) is the predicted pelvis.

In our experiments, we have picked 2D human pose estimation models based on their accuracy, speed and their application in 3D human pose estimation [10, 43, 6] which would in turn, allow us to correctly evaluate and analyze the 3D pose estimators. The 2D models in ascending order of their accuracy (Percentage of Correct Keypoints) are *Mask R-CNN* [21] and Cascaded Pyramid Network (CPN)[8]. Note that these models are substantially different in their methodology and network architectures: while MaskRCNN modifies the original segmentation network to handle pose estimation, CPN uses two different networks (RefineNet and GlobalNet) to perform pose estimation.

As for the 3D human pose estimators, we chose TemporalCNN[43] and Anatomy3D[6]. The training of these 3D models takes place as follows: Both 2D models CPN and Mask RCNN are trained and tested to obtain 2D joints for training and testing sets. These 2D joints are then used as input to the stated 3D pose estimators during the training and testing phases. We trained each model (both on the original and augmented datasets) using the same parameters, given by the authors of the 2D and 3D pose estimation models.

For the Human3.6m dataset, we trained each 2D and 3D model on the subjects S1, S5, S6, S7, S8 and tested on the subjects S9 and S11 for a total of 422420 samples. For the MPI-INF-3DHP dataset, we trained each 2D and 3D models on the subjects S1, S2, S3, S4, S5, S6 and tested on the subjects S7 and S8 for a total of 367522 samples.

In order to boost the computation efficiency of the generation process of the dataset, we save each object using a python dictionary consisting of several features, including area, object image retrieved from the bounding box provided by the dataset, object mask and the category of the object similar to the COCO dataset format.

4.2 2D pose estimation results

In Table 2, we summarize the usefulness of our data augmentation approach when applying both 2D pose estimators under different combinations based on the Human3.6m. The output of all 2D pose estimators includes a layer that predicts a certain number of output heatmaps (probability map of the location of the joint), which is here 17, the same as the number of joints. Under OrgOcc, compared to OrgOrg in Table 2, the produced output for the occluded joint is predicted at a significantly higher distance from the ground truth location when introducing as an occluder object, leading to an increase in the positional error for a specific joint. Table 2 further shows that under OccOcc, i.e., training and then testing the models on the augmented dataset (RealPose) boosts their accuracy in terms of MPJPE error compared to OrgOcc. It is interesting to see that the error under our occlusion dataset (RealPose) is close to that under the original dataset, meaning with our RealPose dataset, the problem of occlusion has been reduced to a large extent.

Since occlusion happens randomly in real-world cases, the effect of training on mixed occluded and not-occluded samples is also important to consider. For this, we

Table 2: Human3.6m: Average MPJPE (in millimeter) of 2D pose estimation models trained and tested on a combination of datasets. OrgOrg indicates training and then testing on the original dataset, OrgOcc training on the original dataset but testing under the occlusion dataset, OccOcc training and testing under the occlusion dataset, MixMix training and testing with the mixed dataset, OccOrg training under occlusion and testing under the original dataset, and MixOrg training under mixed dataset and testing under the original dataset. The gains (in millimeter) of OrgOcc, OccOrg, and MixOrg are with respect to OrgOrg, and gains of MixMix and OccOcc are with respect to OrgOcc.

	MaskRCNN	Gain	CPN	Gain
OrgOrg	11.09	-	9.47	-
OrgOcc	14.64	-3.55	13.00	-3.53
OccOrg	10.52	+0.57	9.41	+0.06
MixOrg	10.46	+0.63	9.35	+0.12
OccOcc	11.70	+2.94	10.87	+2.13
MixMix	10.69	+3.95	9.56	+3.44

created a mixed dataset comprising of both occluded and not-occluded samples (50% from each of RealPose and base dataset), which is the closest to real-world images and scenarios. Table 2 displays the effect of training the two 2D models on MixMix data combinations, the most-realistic case; we see that the models show significant gain compared to OrgOcc. Interesting to note that MaskRCNN shows improved accuracy (of 0.4 millimeter) under MixMix compared to OrgOrg. We also performed simulations under several other data combinations that produced similar results, for example, under MixOcc, i.e., training under mixed set and testing under occlusion set.

We note that MixMix indicates the effect of our data augmentation approach when combined with the original set, resulting in a novel data combination, non-existent in literature. OccOcc entry showcases the effect of training on our RealPose dataset, directly and solely on the occlusion problem which is our main focus when we started our study.

To further show the importance of adding contextual information (i.e., constrained occlusion), we examined the 2D pose estimation models under two combinations:

OccOrg and MixOrg, that is testing under original dataset while training either under occlusion or mixed dataset. As we see in Table 2, the accuracy improves on average by 4% compared to OrgOrg. We associate this improvement with the contextual occlusion which is learned by the model during training on occluded samples.

To examine accuracy per scenario of the Human3.6m datasets, Table 3 displays the MPJPE for each scenario. We see a significant increase in MPJPE between the cases OrgOrg and OrgOcc for each scenario, especially for those in which occlusion is inherently present, such as Discussion and Sitting. With our OccOcc the average estimation error is significantly reduced, with both models MaskRCNN and CPN.

Table 3: Human3.6m scenarios: Comparison of 2D human pose estimators in terms of MPJPE (in millimetres).

	Dir.	Eat.	Dis.	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Average
MaskRCNN - OrgOrg	9.37	10.87	7.47	9.77	12.27	15.27	8.77	8.47	16.47	19.97	11.27	8.17	13.17	6.97	7.17	11.09
MaskRCNN - OrgOcc	10.97	10.94	12.86	13.00	14.26	22.36	11.83	14.56	18.37	20.50	17.23	12.22	21.32	11.61	7.54	14.64
MaskRCNN - OccOcc	5.76	7.98	7.69	11.02	9.19	21.25	7.05	9.31	18.19	19.68	15.25	10.68	19.27	6.77	6.46	11.70
CPN - OrgOrg	8.32	9.15	6.45	8.27	10.25	13.21	7.97	7.85	14.45	15.97	9.27	6.54	11.12	6.48	6.60	9.47
CPN - OrgOcc	13.83	11.68	9.76	12.88	13.40	18.47	10.75	12.38	19.23	18.22	15.65	10.36	11.50	8.27	8.16	13.00
CPN - OccOcc	12.47	10.15	8.54	11.66	11.81	13.24	9.63	10.48	16.02	15.20	13.08	8.20	8.19	7.14	7.38	10.87

Visual comparison based on the Human3.6m test set and the CPN 2D human pose estimator is displayed in Figure 3. By observing this figure, the disfigurement of the skeleton is observed when occlusion is added to the sample. Even in minor cases of occlusion, such as the sample in the second row, the disfigurement is present. OccOcc which is the direct training on an entirely occluded set fixes this issue and reduces the error. Furthermore, by observing the third row of the figure, in which the sample has inherent occlusion due to the special pose, it can be observed that the OccOcc result is closer to the Ground truth than the OrgOrg result, verifying the effectiveness and generalization ability of our approach.

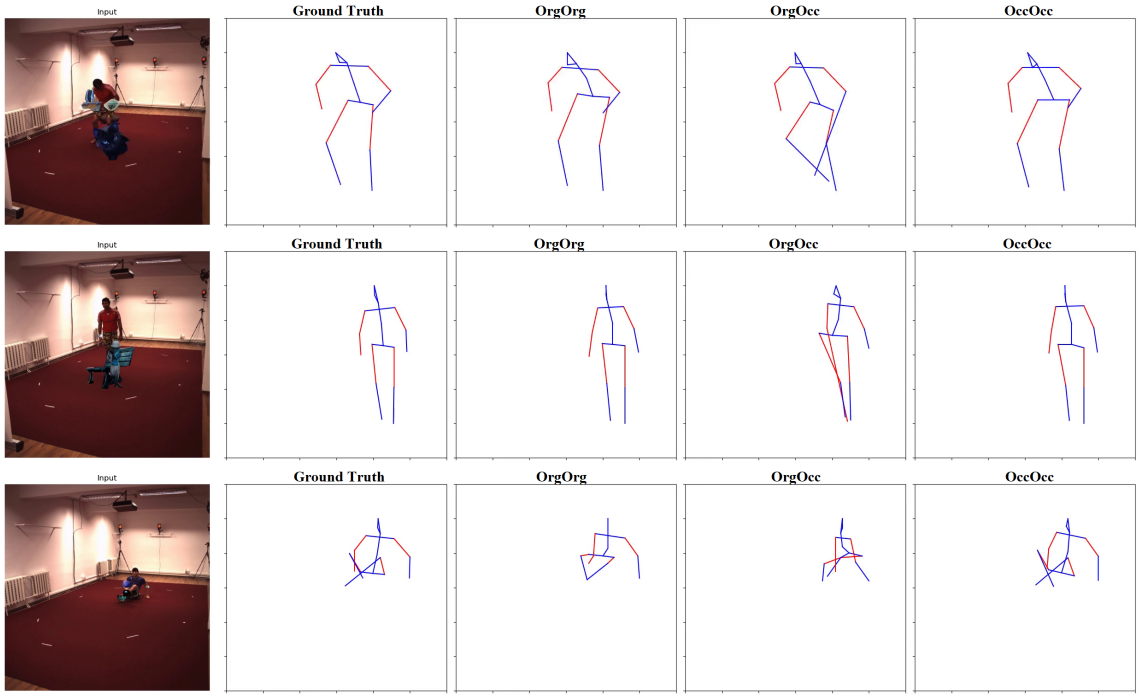

Figure 3: Visual Results of our approach applied to the 2D estimator CPN for different scenarios of the test set of the Human3.6m dataset.

4.3 3D pose estimation results

Table 4 illustrates the results for both 3D pose estimation models on scenarios of the Human3.6m dataset. We see that under OrgOcc the 3D models suffer a major accuracy drop compared to OrgOrg, across all scenarios. Now, when training both models on the RealPose dataset (OccOcc case), a significant accuracy improvement compared to OrgOcc is seen, indicating the effectiveness of our approach on handling the occlusion issue. Comparing 3D results in Table 4 with 2D results in Table 3 for OrgOcc, we notice the significantly higher error of 3D models compared to 2D models. This is because i) 3D pose estimation is a more ill-posed task, compared to 2D estimation because of the depth realization of the final output, creating a list of potential candidates for the final pose; ii) Recent top-performing 3D estimators have a decoupled nature, meaning they estimate based on 2D joints, and not directly regressing 3D points from an RGB image; iii) In our simulations we are training 3D

models on 2D joints extracted from another 2D estimator, and not on the ground truth joints, and thus, we are facing a second layer of error meaning, the 2D joints have an MPJPE themselves, adding up to the 3D MPJPE.

Table 4: Human3.6m scenarios: Comparison of 3D human pose estimators in terms of MPJPE (in millimetres). The CPN 2D model was used to estimate 2D pose for these 3D models.

	Dir.	Eat.	Dis.	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Average
TemporalCNN - OrgOrg	109.64	110.72	69.07	94.88	87.73	110.57	98.56	107.39	91.5	110.51	85.04	99.68	95.14	76.34	70.85	94.5
TemporalCNN - OrgOcc	248.18	252.69	227.02	258.63	233.55	290.37	250.16	272.87	258.32	278.70	229.03	232.13	253.29	251.86	211.66	249.90
TemporalCNN - OccOcc	59.73	63.39	59.54	59.30	69.96	78.01	59.69	56.82	85.76	96.59	65.66	62.66	68.85	52.10	48.12	65.70
Anatomy3D - OrgOrg	102.98	104.84	64.66	88.35	81.96	105.58	89.23	105.12	86.17	105.69	80.63	90.68	90.17	66.93	61.73	88.30
Anatomy3D - OrgOcc	227.13	229.57	208.94	234.48	212.99	269.21	226.81	251.79	241.04	259.14	208.05	211.45	231.45	223.91	187.22	228.20
Anatomy3D - OccOcc	58.04	60.16	57.80	56.99	68.65	77.12	57.90	55.45	83.08	90.44	64.43	60.65	65.11	48.15	43.49	63.20

The error due to occlusion under OrgOcc has blow out of proportions, which is itself an indication of the fragility of the top-performing 3D estimators and the importance of occlusion handling in 3D human pose estimation. Training both 3D models on our RealPose dataset, results in a significant drop in the MPJPE, and we see OccOcc is remarkably better than OrgOrg, indicating the effectiveness of our approach on handling the occlusion issue. This effect is visible across all scenarios, the difficult ones such as Sitting Down and the easiest such as Discussion together.

Table 5 summarizes the effect of applying our method to both 3D estimators under OrgOrg, OrgOcc, OccOcc, OccOrg, MixOrg and MixMix. MixMix is the closest to real-world occlusion; it shows a lower error compared to OccOcc, similar to the 2D case, again indicating the effectiveness of our approach in handling occlusion. Similar to 3D models, OccOrg and MixOrg show how the 3D models benefit from augmented realistic occlusion to even improve accuracy when testing under no occlusion. Visual results for the 3D human pose estimation, using CPN for 2D and TemporalCNN for 3D, based on the OrgOrg, OrgOcc and OccOcc are illustrated in Figure 4. As expected, we observe a heavy distortion when occlusion is added to the samples. As as mentioned before, the error due to the 2D estimator (here, CPN) adds up to the error caused by the 3D estimator (here, TemporalCNN), resulting in the distorted skeleton. In the OccOcc case, we observe the significant improvement in the pose,

Table 5: Human3.6m: Average MPJPE for Anatomy3D and TemporalCNN trained and tested on a combination of datasets (as defined in Table 2). The CPN 2D model was used to estimate 2D pose for these 3D models.

	TemporalCNN	Gain	Anatomy3D	Gain
OrgOrg	94.50	-	88.30	-
OrgOcc	249.90	-155.4	228.20	-139.9
OccOrg	61.70	+32.8	59.50	+28.8
MixOrg	60.90	+33.6	58.20	+30.1
OccOcc	65.70	+184.2	63.20	+165.0
MixMix	61.90	+188.0	60.20	+168.0

when compared to the OrgOrg or the ground truth cases, indicating the effectiveness of our approach starting from the RGB image, to the 2D estimator, ending with the 3D estimator. In some cases, where inherent occlusion is present, such as the third row of Figure 4, the OccOcc result is more accurate than the OrgOrg case.

4.4 MPI-INF-3DHP Results

We applied our approach to the MPI-INF-3DHP dataset in order to test its generalization potential. The pipeline is the same as that of the Human3.6m dataset. The original images are occluded; then the 2D estimators are trained and tested on multiple scenarios of the set. After extracting the 2D joints of the entire dataset using the trained models, the 3D estimators are trained and tested. Results of the 2D pose estimator CPN and the 3D pose estimator TemporalCNN, using 2D poses estimated from CPN, are displayed in Table 6. Overall, all our observations from Human3.6m dataset are confirmed for MPI-INF-3DHP dataset. Specifically, we observe a high increase in MPJPE in the case of OrgOcc, indicating the failure of the models to accurately regress the joints under occlusion. OccOcc case indicates correcting this failure, resulting in a positive gain while the MixMix case shows an even higher gain due to training on the Mixed set.

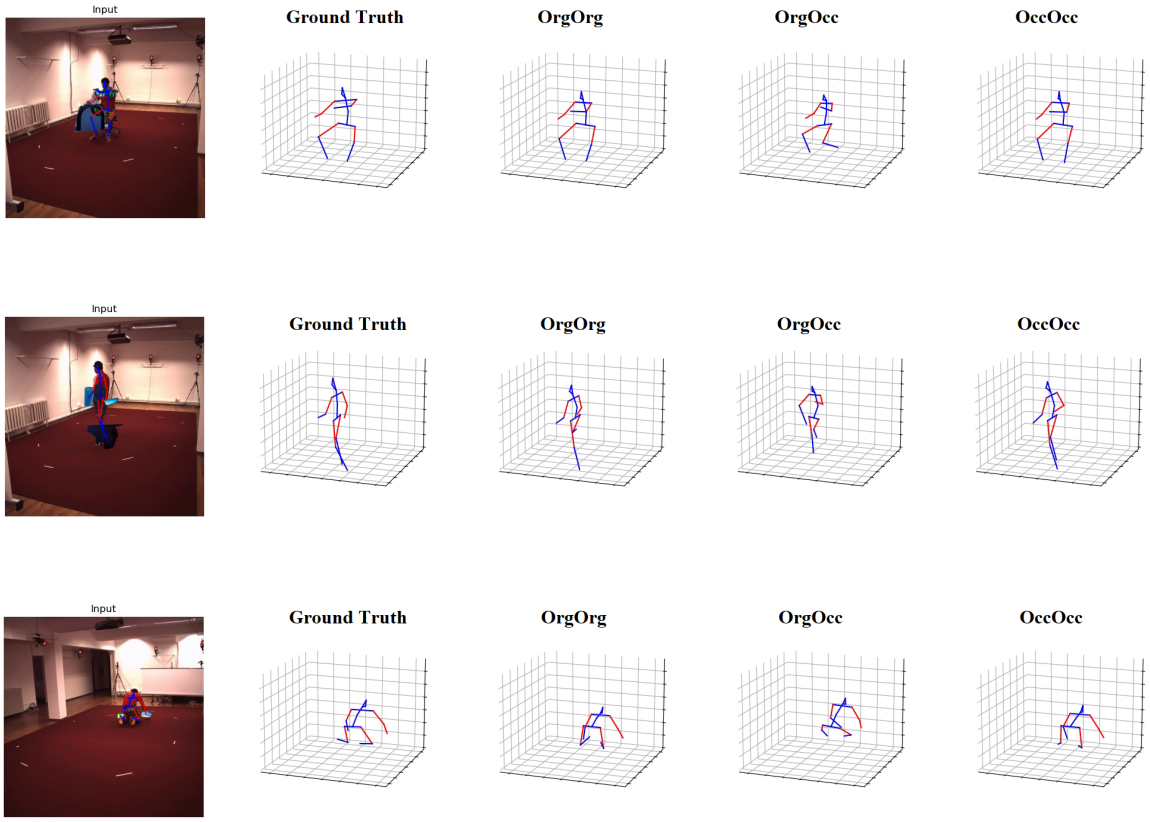

Figure 4: Visual Results of our approach applied to the 3D pose estimator TemporalCNN of the test set of the Human3.6m dataset. In the third row, note the better orientation of the legs under and OccOcc compared to OrgOrg and certainly compared to OrgOcc.

4.5 Discussions

In this work, we explicitly separated different test scenarios so that we would be able to analyze them accurately. This allows for not only evaluating the effectiveness of handling occlusion (OccOcc), but also to make sure it would not fail under the original condition (OccOrg). To our best knowledge, we are the first to benchmark our approach on these multiple levels. Furthermore, the importance of the mixed dataset lies in the fact that related work addressing occlusion by augmentation, apply their methods to the entire training set (Occ case). However, observations of Tables 2,

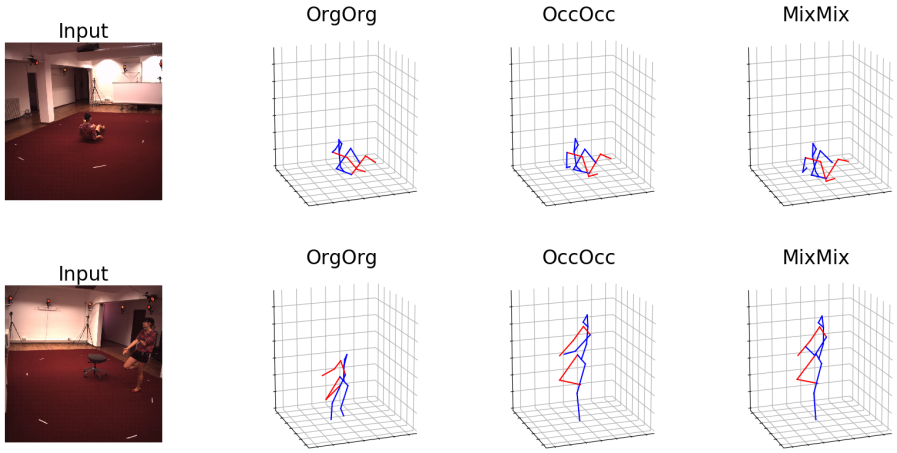

Figure 5: Results of TemporalCNN applied to unlabeled samples with subject performing challenging scenarios. Note the better precision under OccOcc and MixMix in both scenarios compared to OrgOrg: the head is not detected under OrgOrg but is well detected under OccOcc and MixMix. This shows that without proper training under occlusion, contextual information is not properly handled and thus, the model might fail on samples with the slightest occlusion (here a person with a cap).

5, and 6 indicate the effect of considering both occluded and not-occluded samples on boosting the final accuracy of the pose estimation model.

In order to further verify our approach, we tested the 3D models on the unannotated test set of the Human3.6m, namely, subjects 2, 3 and 4 which are typically not used in the literature due to absence of annotations. Samples of these subjects are illustrated in Figure 5, where we used the 3D model TemporalCNN to extract 3D joints from 2D joints estimated using the 2D model CPN.

Results of the 2D pose estimation models clearly indicate the superiority of CPN compared to MaskRCNN. The reason for this better performance is that CPN is inherently designed to be a pose estimation model, while MaskRCNN is an architecture with a variety of applications, namely, object detection, segmentation and of course pose estimation. We apply the same pipeline to the MaskRCNN as well, meaning, training both 3D estimators based on 2D joints, inferenced from MaskRCNN. Our experiments indicated poorer performance of MaskRCNN compared to the CPN model.

Table 6: MPI-INF-3DHP: Average MPJPE for the 2D pose estimator CPN and the 3D pose estimator model TemporalCNN trained and tested on a combination of datasets (as defined in Table 2).

	CPN	Gain	TemporalCNN	Gain
OrgOrg	28.97	-	84.8	-
OrgOcc	82.64	-53.67	218.20	-133.4
OccOrg	13.70	+15.27	57.19	+27.61
MixOrg	13.17	+15.8	56.87	+27.93
OccOcc	16.93	+65.71	65.38	+152.82
MixMix	15.17	+67.47	61.79	+156.41

To show that the proposed joint-occluder category constraints are important, we implemented a method where we augmented any of the 17 joints with any occluder objects out of the 64K objects, randomly cross-matched joints and occluders, and resized the occluder randomly but adaptive to the bounding box of the human subject, meaning, Equation 1 is still utilised. However, the area ratio range [a,b] is now fixed to $a = 0.075$, which is the lowest bound of the area ratios of the constrained augmentation method, and $b = 0.30$ which is the highest bound for the constrained augmentation method. This means that the resizing operation is now independent from the joint-occluder categories. We then trained the 2D pose estimator CPN and the 3D pose estimator TemporalCNN under the revised occlusion dataset. As shown in Table 7, for the OccOcc combination, our method achieves significantly higher accuracy (lower MPJPE) compared with this no-constraint method.

Table 7: Comparison of our method under constrained and not-constrained situations. Results are based on OccOcc, using Human3.6m dataset. CPN 2D pose estimator and TemporalCNN 3D pose estimators were used for evaluation.

OccOcc	2D	3D
no-constraint	13.37mm	80.4mm
constrained	10.87mm	65.7mm

Chapter 5

Conclusion and Future Work

5.1 Conclusion

In this thesis, we studied the effect of occlusion on the task of human pose estimation using CNN. We observed a significant decrease of accuracy of both 2D and 3D pose estimation models under occlusion. We then proposed an approach to augment base datasets with occlusion data (called occluders). We have chosen different occluder levels categories and limited each to a particular area of the human body where it would more probably appear, and utilized morphological erosion operation to blend the occluder onto the original image. We have then applied top-performing 2D and 3D human pose estimators which are CNN based, and tested them on a combination of occluded, not-occluded and mixed samples. The combinations are base dataset (no occlusion added), occluded dataset (the base dataset augmented with occlusion), and mixed dataset (50% from the base dataset and 50% from the augmented dataset). We have divided each dataset into training and testing sets. Then we first trained the CNN models on the training set and then tested on the test dataset of the respective combination. The results of our extensive experiments have indicated that while without proper training, occlusion lowers the accuracy of the CNN model; on the

other side, using the augmented samples for training will incorporate occlusion into the model and as a result significantly boosts their accuracy. Our approach resulted in creation of different augmented datasets: OccRealPose (consisting of occluded samples only) and MixRealPose (consisting of occluded and not-occluded samples). A significant outcome of our work is that, our method results in a higher accuracy of CNN models when trained under occluded samples but tested under the original (not-occluded) samples, indicating the model achieves a higher understanding of the dependency of different joints induced to it. Our proposed approach is independent of the base dataset and can be used on any human pose estimation dataset and network. These datasets as well as the code of our approach are available for download under https://users.encs.concordia.ca/∼amer/RealPose/. Our approach has been tested for single-person pose estimation, however, it can be extended to multi-person pose estimation by providing a bounding box for each human subject and and occluding each in an isolated manner.

5.2 Future Work

While our occlusion augmentation method provides an excellent opportunity to directly address occlusion, it still lacks occlusion labels, meaning the exact location of the occurring occlusion. This opens the way to two main future research opportunities. First, to design and embed occlusion labels to the occluded image both for 2D and 3D human pose estimation. Second, a model can be designed to actually train on the occlusion samples, meaning, taking into account the occlusion labels and enforce a penalty, mark the sample as a difficult sample, or work as a pose discriminator.

Regarding the dataset itself, while we did our best efforts to pay special attention to the reality of the dataset it might still contain artifacts from the blending and resizing resulting in blurry samples. Recent advancements in Generative Adversarial

Networks indicate that they can be used to create conditional samples, meaning we specify the details of the output image. The output samples are realistic to a degree that it is undetectable by the human eye.

Bibliography

[1] Amin Ansarian and Maria Amer. Realistic augmentation for effective 2D human pose estimation under occlusion. In *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2021 (accepted).

[2] Yanrui Bin, Xuan Cao, Xinya Chen, Yanhao Ge, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, Changxin Gao, and Nong Sang. Adversarial semantic data augmentation for human pose estimation. In *European Conference on Computer Vision*, pages 606–622. Springer, 2020.

[3] Yujun Cai, Liuhao Ge, Jun Liu, Jianfei Cai, Tat-Jen Cham, Junsong Yuan, and Nadia Magnenat Thalmann. Exploiting spatial-temporal relationships for 3D pose estimation via graph convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2272–2281, 2019.

[4] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: realtime multi-person 2D pose estimation using part affinity fields. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):172–186, 2019.

[5] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2D pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017.

[6] Tianlang Chen, Chen Fang, Xiaohui Shen, Yiheng Zhu, Zhili Chen, and Jiebo Luo. Anatomy-aware 3D human pose estimation in videos. *arXiv preprint arXiv:2002.10322*, 2020.

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

[8] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7103–7112, 2018.

[9] Yu Cheng, Bo Yang, Bo Wang, and Robby Tan. 3D human pose estimation using spatio-temporal networks with explicit occlusion training. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:10631–10638, 04 2020.

[10] Yu Cheng, Bo Yang, Bo Wang, and Robby T Tan. 3D human pose estimation using spatio-temporal networks with explicit occlusion training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10631–10638, 2020.

[11] Yu Cheng, Bo Yang, Bo Wang, Wending Yan, and Robby T Tan. Occlusion-aware networks for 3D human pose estimation in video. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 723–732, 2019.

[12] Xiao Chu, Wei Yang, Wanli Ouyang, Cheng Ma, Alan L Yuille, and Xiaogang Wang. Multi-context attention for human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1831–1840, 2017.

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[14] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

[15] Nikita Dvornik, Julien Mairal, and Cordelia Schmid. Modeling visual context is key to augmenting object detection datasets. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 364–380, 2018.

[16] Debidatta Dwibedi, Ishan Misra, and Martial Hebert. Cut, paste and learn: Surprisingly easy synthesis for instance detection. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[17] Bin et al. Adversarial semantic data augmentation for human pose estimation. In *European Conference on Computer Vision (ECCV)*, pages 606–622, 2020.

[18] Munea et al. The progress of human pose estimation: A survey and taxonomy of models applied in 2D human pose estimation. *IEEE Access*, 8:133330–133348, 2020.

[19] Peng et al. Jointly optimize data augmentation and network training: Adversarial data augmentation in human pose estimation. In *Conference on Computer Vision and Pattern Recognition*, pages 2226–2234, 2018.

[20] Golnaz Ghiasi, Yi Yang, Deva Ramanan, and Charless C Fowlkes. Parsing occluded people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2401–2408, 2014.

[21] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[23] Yihui He, Rui Yan, Katerina Fragkiadaki, and Shoou-I Yu. Epipolar transformers. *arXiv*, pages arXiv–2005, 2020.

[24] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9:1735–80, 12 1997.

[25] Fuyang Huang, Ailing Zeng, Minhao Liu, Qiuxia Lai, and Qiang Xu. Deepfuse: An IMU-aware network for real-time 3D human pose estimation from multi-view image. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 429–438, 2020.

[26] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3D human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, 2014.

[27] Karim Iskakov, Egor Burkov, Victor Lempitsky, and Yury Malkov. Learnable triangulation of human pose. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7718–7727, 2019.

[28] Karim Iskakov, Egor Burkov, Victor Lempitsky, and Yury Malkov. Learnable triangulation of human pose. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7718–7727, 2019.

[29] Adam Kortylewski, Qing Liu, Huiyu Wang, Zhishuai Zhang, and Alan Yuille. Combining compositional models and deep networks for robust object classification under occlusion. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1333–1341, 2020.

[30] Shichao Li, Lei Ke, Kevin Pratama, Yu-Wing Tai, Chi-Keung Tang, and Kwang-Ting Cheng. Cascaded deep monocular 3D human pose estimation with evolutionary training data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[31] Wenbo Li, Zhicheng Wang, Binyi Yin, Qixiang Peng, Yuming Du, Tianzi Xiao, Gang Yu, Hongtao Lu, Yichen Wei, and Jian Sun. Rethinking on multi-stage networks for human pose estimation. *arXiv preprint arXiv:1901.00148*, 2019.

[32] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[33] W. Liu, Q. Bao, Y. Sun, and T. Mei. Recent advances in monocular 2D and 3D human pose estimation: A deep learning perspective, 2021.

[34] C. Luo, H. Mobahi, and S. Bengio. Data augmentation via structured adversarial perturbations, 2020.

[35] William J. McNally, Kanav Vats, Alexander Wong, and John McPhee. Evopose2D: Pushing the boundaries of 2D human pose estimation using neuroevolution. *CoRR*, abs/2011.08446, 2020.

[36] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3D human pose estimation

in the wild using improved cnn supervision. In *3D Vision (3DV), 2017 Fifth International Conference on*. IEEE, 2017.

[37] Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Posefix: Model-agnostic general human pose refinement network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7773–7781, 2019.

[38] Francesc Moreno-Noguer. 3D human pose estimation from a single image via distance matrix regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2823–2832, 2017.

[39] Alejandro Newell, Zhiao Huang, and Jia Deng. Associative embedding: End-to-end learning for joint detection and grouping. *arXiv preprint arXiv:1611.05424*, 2016.

[40] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, pages 483–499. Springer, 2016.

[41] Andrew Ng. A chat with Andrew on MLOps: From model-centric to data-centric AI.

[42] Curtis Northcutt, Lu Jiang, and Isaac Chuang. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, 70:1373–1411, 2021.

[43] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3D human pose estimation in video with temporal convolutions and semi-supervised training. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, pages 7753–7762, 2019.

[44] Haibo Qiu, Chunyu Wang, Jingdong Wang, Naiyan Wang, and Wenjun Zeng. Cross view fusion for 3D human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4342–4351, 2019.

[45] István Sárándi, Timm Linder, Kai O Arras, and Bastian Leibe. How robust is 3D human pose estimation to occlusion? In *IROS Workshop - Robotic Co-workers 4.0*, 2018.

[46] Jon Shlens, Ekin Dogus Cubuk, Quoc Le, Tsung-Yi Lin, Barret Zoph, and Golnaz Ghiasi. Learning data augmentation strategies for object detection, November 21 2019. US Patent App. 16/416,848.

[47] Leonid Sigal, Alexandru O Balan, and Michael J Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International journal of computer vision*, 87(1-2):4, 2010.

[48] Leonid Sigal and Michael J Black. Measure locally, reason globally: Occlusion-sensitive articulated pose estimation. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 2041–2048. IEEE, 2006.

[49] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5693–5703, 2019.

[50] Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3D human pose in the wild using imus and a moving camera. In *European Conference on Computer Vision (ECCV)*, sep 2018.

[51] S. Vosoughi and M. A. Amer. Deep 3D human pose estimation under partial body presence. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 569–573, 2018.

[52] Jacob Walker, Kenneth Marino, Abhinav Gupta, and Martial Hebert. The pose knows: Video forecasting by generating pose futures. In *Proceedings of the IEEE international conference on computer vision*, pages 3332–3341, 2017.

[53] Bastian Wandt and Bodo Rosenhahn. Repnet: Weakly supervised training of an adversarial reprojection network for 3D human pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7782–7791, 2019.

[54] Ch. Wang, F. Zhang, and Sh. Sam Ge. A comprehensive survey on 2D multi-person pose estimation methods. *Engineering Applications of Artificial Intelligence*, 102, 2021.

[55] Keze Wang, Liang Lin, Chenhan Jiang, Chen Qian, and Pengxu Wei. 3D human pose machines with self-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 42(5):1069–1082, 2019.

[56] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. `https://github.com/facebookresearch/detectron2`, 2019.

[57] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *European Conference on Computer Vision (ECCV)*, 2018.

[58] Yuanlu Xu, Wenguan Wang, Xiaobai Liu, Jianwen Xie, and Song-Chun Zhu. Learning pose grammar for monocular 3D pose estimation, 2019.

[59] Zongxin Yang, Xin Yu, and Yi Yang. Dsc-posenet: Learning 6dof object pose estimation via dual-scale consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3907–3916, 2021.

[60] Feng Zhang, Xiatian Zhu, Hanbin Dai, Mao Ye, and Ce Zhu. Distribution-aware coordinate representation for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7093–7102, 2020.

[61] T. Zhang, J. Wang, Q. Zhu, and B. Yin. See through occlusions: Detailed human shape estimation from a single image with occlusions. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 2646–2650, 2020.

[62] Long Zhao, Xi Peng, Yu Tian, Mubbasir Kapadia, and Dimitris N Metaxas. Semantic graph convolutional networks for 3D human pose regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3425–3435, 2019.

[63] Jian Zheng, Yifan Wang, Xiaonan Zhang, and Xiaohua Li. Classification of severely occluded image sequences via convolutional recurrent neural networks. In *2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 395–399. IEEE, 2018.

[64] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *AAAI*, pages 13001–13008, 2020.