Learning to Train with Synthetic Humans

David T. Hoffmann[1], Dimitrios Tzionas[1], Michael J. Black[1], and Siyu Tang[1,2,3]

[1] Max Planck Institute for Intelligent Systems, Germany
[2] University of Tübingen, Germany [3] ETH Zürich
dhoffmann, dtzionas, black, stang@tuebingen.mpg.de

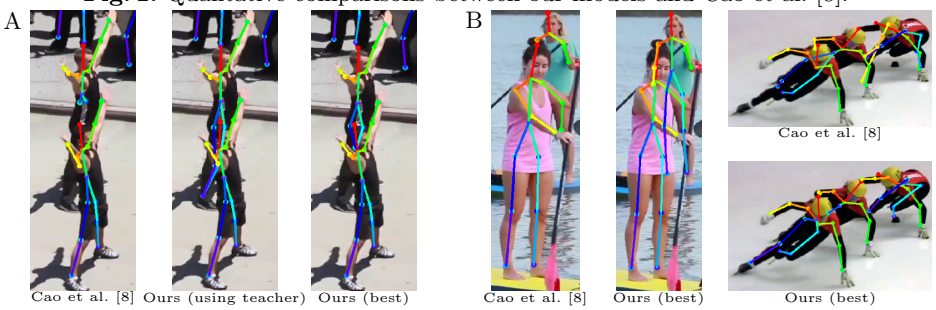
Abstract. Neural networks need big annotated datasets for training. However, manual annotation can be too expensive or even unfeasible for certain tasks, like multi-person 2D pose estimation with severe occlusions. A remedy for this is synthetic data with perfect ground truth. Here we explore two variations of synthetic data for this challenging problem; a dataset with purely synthetic humans and a real dataset augmented with synthetic humans. We then study which approach better generalizes to real data, as well as the influence of virtual humans in the training loss. Using the augmented dataset, without considering synthetic humans in the loss, leads to the best results. We observe that not all synthetic samples are equally informative for training, while the informative samples are different for each training stage. To exploit this observation, we employ an adversarial student-teacher framework; the teacher improves the student by providing the hardest samples for its current state as a challenge. Experiments show that the student-teacher framework outperforms normal training on the purely synthetic dataset.

1 Introduction

The broad success of deep neural networks comes at a price: the ever growing need for huge amounts of labeled training data. For many tasks, the lack of data seems to be one of the major limiting factors of progress. It is particularly problematic for the tasks where manual labeling requires significant human effort, or is even unfeasible. For example, in multi-person 2D pose estimation, a major challenge is that people are often partially visible. Manual annotation of body joints that are severely occluded is error prone and the resulting labels are noisy. Computer graphics can help to resolve these issues. 3D rendering engines offer the opportunity to generate a large amount of data with perfect labels: e.g., the location of occluded body parts and the precise pose of the camera.

Nowadays, large scale synthetic datasets with reasonable realism can be generated relatively easy and the idea of synthesizing training data has been widely explored. In general, there are two common strategies for generating a synthetic dataset: rendering a purely synthetic dataset, and augmenting real training images with synthetic instances. The advantage of the former is the full control over the virtual 3D world and ability to generate high variance datasets [4,5,12,29,35,36,38,44]. The advantage of the second approach is that some of the instances in the dataset are real, resulting in overall higher realism [2,11,40].

Fig. 1. Qualitative comparisons between our models and Cao et al. [8].



We generate both types of synthetic datasets. One purely synthetic dataset and a mixed dataset, which is generated by augmenting the MPII pose estimation dataset [3] with synthetic humans. In particular, we design these datasets to improve on frequent failure cases that we observe with state-of-the-art models (see Fig 1), namely uncommon camera angles and strong occlusion. By comparing generalization performance using these two datasets we obtain insights into which way of generating data is preferable. We further investigate how strongly the lack of photorealism of the synthetic humans limits generalization. To make the synthetic data more realistic, we propose a simple synthetic-to-real human style transfer algorithm, based on the work of Dundar et al. [10].

These experiments show that naive training with synthetic data leads only to limited improvements. One explanation is that training on large synthetic datasets leads to overfitting of the model to the features of synthetic data. We observe that some synthetic images convey more information than others. Overfitting to features of synthetic data could be limited by generating only useful, i.e. difficult synthetic data, and thus limiting the training on synthetic data.

As a step in this direction, we propose a method to use synthetic datasets more effectively. Specifically, we introduce an adversarial student-teacher framework. The teacher learns online which training data is still difficult. This information is then used to increase the sampling probability of similar examples. By taking into account feedback from the student, the teacher keeps on updating the sampling probabilities throughout training and adapts them to the specific needs of the student. Training with the teacher on the purely synthetic data outperforms normal training.

Our contributions can be summarized as follows: 1) We propose a large-scale synthetic multi-person dataset, a mixed dataset, and a domain-adapted version of the latter. 2) We explore which way of generating synthetic data is superior for our task. 3) We propose a student-teacher framework to train on the most difficult images and show that this method outperforms random sampling of training data on the synthetic dataset. We provide datasets and code[1].

[1] `https://ltsh.is.tue.mpg.de`

2 Related Work

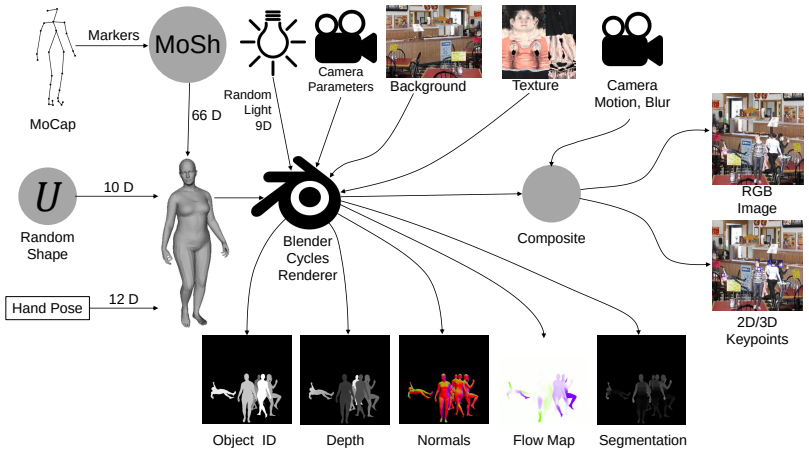
Synthetic datasets with humans. The need for labelled training data has fueled development of datasets with synthetic humans. Many methods use 3D models of the human body to generate data [4,5,16,35]. Other approaches augment 3D training data by utilizing 2D pose datasets [9,37], while [40,43] augment datasets with cut-outs of objects or animals. Closely related to our approach, [36,44] render the SMPL model [27] on top of random indoor images. These methods generate datasets with a single synthetic human. Multi-person datasets were created by employing video games for pedestrian detection [29] and pose tracking [12]. Similarly, [30] develop a simulation environment in a game engine, including virtual humans. Related to our approach, [38] augment real training images similar to the ones in [44] but with multiple synthetic humans occluding each other.

Domain Adaptation. The quality of synthetic data is often insufficient to generalize well to real data. Several domain adaptation methods have been developed to overcome this problem. Shrivastava et al. [42] train a Generative Adversarial Network (GAN) to refine synthetic images, while keeping the label information intact. Recently, Cycle-GAN has been used to map images from one domain to another [4,31]. However, GAN-based methods are prone to unstable training and require tedious hyper-parameter tuning. As a practical alternative, recently [10] proposed domain stylization to stylize synthetic images to real ones, using the fast photorealistic image stylization method of [25].

Human pose estimation. Multi-person pose estimation has attracted substantial attention over the last years [8,14,15,19,23,28,32,33]. One of the most popular datasets is the MPII multi-person pose estimation dataset [3]. Among the best performing methods on MPII are: [33], which uses a pose partition network, [28], which uses context information, [15], which refines pose predictions, and [32], which predicts "tag maps" to solve the grouping problem. The most widely used method is OpenPose [8], a bottom-up approach that first predicts keypoints and then estimates Part Affinity Fields (PAFs) to group them.

Learning to train. Bengio et al. [6] introduce curriculum learning for learning systems, exploiting the idea that different data samples are informative at different training stages. As a proof of concept they manually define the samples for each stage with gradually increasing difficulty. Multiple methods have focused on automating curriculum learning [7,13,20,24]. These approaches try to maximize information gain during training by carefully monitoring the learning success of the model. Alternatively, adversarial methods [21,22,41] try to pick the hardest samples at each training stage; [21] favors samples resulting in higher loss, [22] learns weights for the loss of each training sample as a soft curriculum, while [41] uses online hard example mining for object detection. Peng et al. [34] propose an adversarial training scheme to optimize data augmentation online. They train two teacher networks to learn a probability distribution over the hyper-parameters for data augmentation; one predicts the most difficult image rotations, while the other predicts parameters for deep feature occlusion.

Fig. 2. Schematic of the data generation pipeline.



3 Multi-person Synthetic Data

In the following we describe the generation of the datasets (Sections 3.1, 3.2, 3.3) and the domain adaptation used to increase visual appearance of the mixed dataset (Section 3.4).

3.1 Data Generation Pipeline

To generate realistic synthetic training data we build on top of [36,44]. We use multiple different sources of data to build a realistic synthetic scene. Images and ground truth annotations are rendered using Cycles, the rendering engine of Blender[2]. An overview can be seen in Fig. 2.

Body Model and MoCap Data. We use the parametric body model SMPL+H [39] to generate realistic synthetic humans. SMPL+H is parameterized by pose $\theta \in \mathbb{R}^{78}$ and shape $\beta \in \mathbb{R}^{10}$. We collect realistic pose and shape parameters by fitting SMPL+H to standard MoCap data by using MoSh [26].

Details. We draw inspiration from [36,44] who generate small video sequences, each having different but fixed parameters for the position, pose, shape and texture of a synthetic human, the background image, camera position, lighting, etc. In contrast, we generate single images and render multiple synthetic humans, while randomizing the number of them. As a result, we generate a dataset with much higher variance. Images with inter-penetrating meshes of virtual humans are rejected to avoid artifacts in the generated ground truth.

Further details regarding the description of the data generation pipeline, posing of hands and a quantitative comparison of our datasets to other datasets can be found in **Supp. Mat.**

[2] `https://www.blender.org`

Fig. 3. Example images from the purely synthetic dataset. It contains high occlusion, extreme poses, various camera angles and various challenging backgrounds.

3.2 Synthetic Dataset

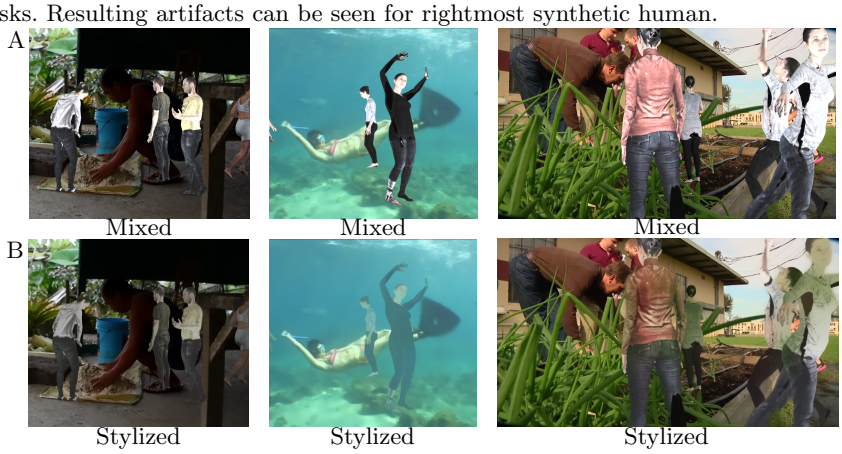
Background Images. To generalize well to in-the-wild pose estimation datasets the background images should come from many different scenes. To this end we use images from SUN397 [45] and reject all images with a resolution smaller than 512×512 pixels to ensure high quality backgrounds. Additionally, we reject all images containing humans, as we do not have ground truth annotations for them. We use mask-RCNN [1,17] as our human detector.

Generative Factors. We sample the number of synthetic humans per image from a Poisson distribution with $\lambda = 9$, to encourage many humans per image, while avoiding too extreme values. The datasets of [36,44] have only very small variance in camera position. However, preliminary experiments show that the camera position significantly influences the difficulty of multi-person pose estimation. Therefore we increase the range of possible camera positions, by sampling the camera pitch uniformly from $[0, 45]^\circ$. The resolution of the final rendered images is set to 640×640 pixels. We refer to this dataset as $\mathcal{D}_S$.

3.3 Mixed Dataset

We build upon the finding of [40] that realistic occluding objects lead to larger improvements than abstract objects. We choose our occluders to be from the same class as our target objects; i.e. humans, to simulate crowded scenes with multiple humans. To generate the dataset we use the pipeline described above with a few differences. Instead of SUN397 [45] we use the training images of the MPII human pose dataset [3] as background images. To keep the MPII ground truth intact, we render the images with the same resolution as the background MPII image, and keep the camera pose fixed. We then augment the MPII human pose dataset by superimposing synthetic humans. Their number is drawn from a Poisson distribution with $\lambda = 4$ to introduce interesting and intense occlusions as shown in Fig. 4 (A), without extreme occlusions by too many synthetic humans. We render each of the $15,956$ images in our training set 5 times with different parameters for increased variance. We refer to this dataset as $\mathcal{D}_M$.

Fig. 4. (A) Example images from $\mathcal{D}_M$. (B) Corresponding images of $\mathcal{D}_{Style}$. For the last image, the segmentation network included non-human parts in the segmentation masks. Resulting artifacts can be seen for rightmost synthetic human.



3.4 Domain Stylization

The appearance of real and synthetic humans differs strongly. Factors contributing to these differences are the low quality of textures and differences in lighting conditions for synthetic humans and background images. Additionally, the small number of human textures limits the variability. These differences in appearance might limit the generalization. We draw inspiration from Dundar et al. [10] to reduce these differences by using the fast photorealistic style transfer method of [25]. Style transfer methods require a pair of images as input, a content image I_C and a style image I_S. While the style of these images can be largely different, their content should have similarities. Finding such pairs is a non-trivial problem. However, for the $\mathcal{D}_M$, we have a canonical choice of image pairs: the image from $\mathcal{D}_M$ and its background as I_C and I_S, respectively.

Naive application of style transfer methods on the whole image, leads to severe artifacts. Therefore, only the style of semantically similar classes should be transferred. To obtain a good semantic segmentation network Dundar et al. [10] iterate between stylizing a dataset and training a network for semantic segmentation on the stylized dataset. Here, we are only interested in transferring the style of real to synthetic humans. Fine-grained human detection is important to avoid parts of the background bleeding into the foreground after style transfer (see Fig. 4 (B) right panel). We therefore employ Mask-RCNN [1,17] to predict pixel-wise masks for humans. Ground truth masks for the synthetic humans are generated during data generation. Since the style-transfer algorithm can not handle images of arbitrary size, we rescale the larger images to 600 pixels before applying the style transfer. We refer to the resulting dataset as $\mathcal{D}_{Style}$. Examples can be seen in Fig. 4 (B).

4 Learning to Train with Multi-person Synthetic Data

We now turn to the task of training a pose estimation network with synthetic data. We first provide general information about our model. This is followed by a brief description of the training procedure. Finally, we detail the student-teacher framework and explain how the teacher is trained.

Pose Estimation Network. For all experiments we use our Tensorflow implementation of the network proposed in [8], which we will refer to as *OpenPose* network. Our training differs only slightly from [8]; details are provided in **Supp. Mat.** Because of these differences, we use a self-trained model on real data as the baseline throughout the paper.

Training with Synthetic Data. Following the advice of [18], we freeze the weights of our feature extractor whenever training on synthetic data. In particular, we freeze the first 4 layers of the OpenPose network. Additionally, we make sure that each batch is composed of 50% real and 50% synthetic images. More details on our training procedure with synthetic data are provided in **Supp. Mat.**

Grouping. Sampling the most difficult samples with higher probability comes with the risk of oversampling a small amount of data. To avoid such behavior, and ease the task for the teacher, we group the synthetic data into meaningful groups. We found empirically that the position of the camera and the distance of people in an image contribute to the difficulty of multi-person pose estimation. Thus, we use these two image characteristics to group our data. We assume 10 *groups* to yield a good trade-off between precision and difficulty for the teacher. For minimal distance grouping[3], denoted as mD, we decide for linearly spaced values between $[0, 640)$ px. For the camera pitch grouping, denoted as C we space the group boundaries linearly in the interval $[\min(X)+\mathrm{Var}(X),\ \max(X)-\mathrm{Var}(X))$, where X contains all values for the camera pitch in the dataset.

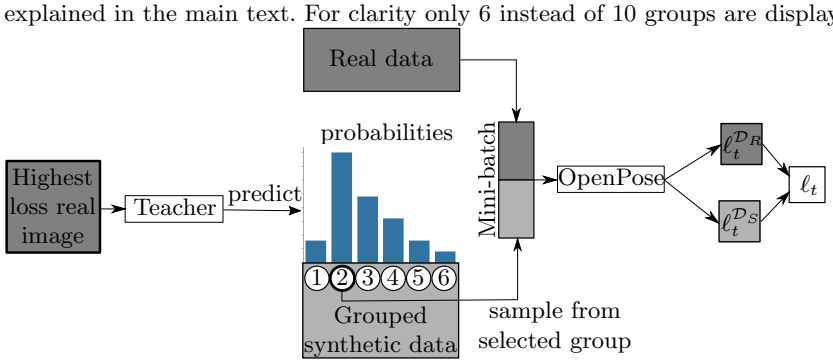
4.1 Adversarial Teacher

The teacher network is trained simultaneously with the student network to adapt the sampling strategy dynamically to the current training state of the student. A schematic of the forward pass can be seen in Fig. 5. The input to the teacher should represent the training state of the student. We choose the real image resulting in highest mean loss per joint within the previous N training steps. This provides some information about the type of images that are still difficult for the student. The output of the teacher network is a probability distribution $\tilde{\boldsymbol{P}}$ over a set of groups $\{g_1, \ldots, g_i\}$. This probability distribution is used to sample one of them. For the next N training steps, synthetic training samples are drawn from this group only.

Training the Teacher. The objective of the adversarial teacher should be to maximize the loss, $\ell_t^{\mathcal{D}_S}$. Unfortunately, sampling and data augmentation

[3] We sample persons not images. The image is cropped around this person. The minimal distance is defined as smallest distance of this person to any other person.

Fig. 5. Diagram of the forward pass. The total loss is denoted as ℓ_t, the loss on real images as $\ell_t^{\mathcal{D}_R}$ and on synthetic images as $\ell_t^{\mathcal{D}_S}$. The reward signal and backward pass are explained in the main text. For clarity only 6 instead of 10 groups are displayed.



are non-differentiable operations, prohibiting end-to-end training of the teacher network. An alternative method to provide a supervision signal is needed. We draw inspiration from [34] and employ a reward/penalty training scheme. To determine whether the teacher is rewarded or penalized we monitor $\ell_t^{\mathcal{D}_S}$. If $\ell_t^{\mathcal{D}_S} \geq \ell_{t-1}^{\mathcal{D}_S}$ the teacher succeeded in finding more difficult examples than before and is rewarded. Unfortunately, $\ell_t^{\mathcal{D}_S}$ has a high variance. To reduce the variance of the teaching signal we reward if

$$\ell_t^{\mathcal{D}_S} \geq \frac{1}{H} \sum_{h=0}^{H} \ell_{t-1-h}^{\mathcal{D}_S}, \tag{1}$$

where H denotes the number of past loss values to be considered. To avoid favoring images with many people, we use the mean loss per joint on the image. Eq. 1 provides the direction of the gradient descent step but no ground truth is given to compute the gradients. To efficiently get a reward signal we follow [34] and increase the probability of a group being chosen, if the teacher gets rewarded. Probabilities for other groups are decreased accordingly. Formally, we update P_i and P_j, where i denotes the selected group and $j \neq i$ denotes all other groups, by

$$P_i = \tilde{P}_i + \delta\alpha\tilde{P}_i, \text{ and } P_j = \tilde{P}_j - \delta\frac{\alpha\tilde{P}_i}{|g|-1}. \tag{2}$$

Here $\tilde{P}_i$ denotes the prediction of the teacher, P_i is the updated pseudo ground truth probability, $0 \leq \alpha \leq 1$ controls the size of the update, $|g|$ denotes the number of groups and δ is a sign indicator $\delta = \{+1, \text{ if Eq. 1 holds; } -1, \text{ otherwise}\}$. Finally, we obtain gradients to update the teacher network by computing the KL-divergence loss between $\boldsymbol{P}$ and $\tilde{\boldsymbol{P}}$. Information on optimization related hyperparameters and the architecture are provided in **Supp. Mat.**

During training we face a exploration/exploitation trade-off. The group with highest probability, might not be optimal. To overcome this problem, we sample

Table 1. Results on the held out validation set. $\mathcal{M}_{\mathcal{D}_R}$ is trained only with real data. $\mathcal{M}_{\mathcal{D}_S}$ was trained solely on synthetic data. $\mathcal{M}_{\mathcal{D}_R+\mathcal{D}_S}$ is trained on real and synthetic data. $\mathcal{M}_{\mathcal{D}_R+\mathcal{D}_M}$ was trained on real and the mixed dataset.

Model	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	mAP
$\mathcal{M}_{\mathcal{D}_R}$	**91.3**	89.1	79.2	70.4	**75.9**	71.5	66.7	77.7
$\mathcal{M}_{\mathcal{D}_S}$	37.9	23.5	12.7	7.3	5.6	3.4	3.2	13.4
$\mathcal{M}_{\mathcal{D}_R+\mathcal{D}_S}$	91.1	89.2	80.5	71.0	75.2	**73.6**	**68.1**	**78.4**
$\mathcal{M}_{\mathcal{D}_R+\mathcal{D}_M}$	**91.3**	**89.5**	**80.7**	**71.7**	75.4	72.5	67.7	**78.4**

a group from a uniform distribution instead of the predicted probabilities with probability $\epsilon = 0.1$.

5 Experiments

We test our models on the MPII multi-person pose dataset [3]. For that purpose we use the toolkit provided by [3]. Following the standard validation procedure, the test metric is only computed for people within close proximity. We split the real training data, denoted as $\mathcal{D}_R$, into a training and a validation set. Our validation set consist of 343 randomly selected groups of people in close proximity. The respective images are not used for training. We report mAP (mean Average Precision), the main metric of the benchmark [3] for each model.

5.1 Which Dataset Generalizes Best?

We hypothesize that multi-person pose estimation methods are limited by a lack of training data. To test this hypothesis we train our model on $\mathcal{D}_S$ and $\mathcal{D}_R$+$\mathcal{D}_S$. Interestingly, for $\mathcal{M}_{\mathcal{D}_S}$, the model resulting from training only on $\mathcal{D}_S$, mAP is very low, suggesting that the model overfits to the synthetic data. However, when training on $\mathcal{D}_R$+$\mathcal{D}_S$ the mAP improves over the mAP of $\mathcal{M}_{\mathcal{D}_R}$ (Tab. 1).

While synthetic data can improve the accuracy of multi-person pose estimation, the improvements are relatively small given the extensive amount of additional training data. Multiple factors might limit the generalization. It could well be that the dataset bias between the synthetic and real dataset is just too strong. Generating the $\mathcal{D}_M$ is a straightforward way of generating a dataset with similar dataset bias as the original dataset. By training on $\mathcal{D}_M$ we can quantify the influence of it on the generalization. When training with $\mathcal{D}_M$ we consider only real humans as samples. The rationale behind that decision is that we primarily want to increase the frequency of occlusion of real humans. However, the network is also trained on all synthetic humans that are within the cropped training image. As can be seen in Tab. 1, training on $\mathcal{D}_R$+$\mathcal{D}_M$ results in similar accuracy as $\mathcal{M}_{\mathcal{D}_R+\mathcal{D}_S}$. Therefore, the two methods of generating data are equivalently good. The dataset bias seems not to be the main limiting factor.

Table 2. Results for the models $\mathcal{M}_{\mathcal{D}_R+\mathcal{D}_M}$, $\mathcal{M}_{\mathcal{D}_R+\mathcal{D}_{Style}}$, $\mathcal{M}_{\mathcal{D}_R+\mathcal{D}_M+\text{masks}}$ and $\mathcal{M}_{\mathcal{D}_R+\mathcal{D}_{Style}+\text{masks}}$ datasets, where "+ masks" denotes masking out loss generated by synthetic humans. All reported results are on the held out validation set.

Model	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	mAP
$\mathcal{M}_{\mathcal{D}_R+\mathcal{D}_M}$	91.3	89.5	80.7	71.7	75.4	**72.5**	67.7	78.4
$\mathcal{M}_{\mathcal{D}_R+\mathcal{D}_M+\text{masks}}$	**92.3**	**90.9**	80.5	**72.2**	76.0	71.7	68.3	78.9
$\mathcal{M}_{\mathcal{D}_R+\mathcal{D}_{Style}}$	91.8	89.8	80.4	70.9	75.5	71.6	67.9	78.3
$\mathcal{M}_{\mathcal{D}_R+\mathcal{D}_{Style}+\text{masks}}$	91.6	90.6	**80.8**	71.8	**77.7**	72.2	**68.8**	**79.1**

5.2 Does Stylization Improve Generalization?

The generalization might be limited by the appearance of synthetic humans. To overcome the limited generalization and measure how the difference in appearance influences performance we train on $\mathcal{D}_R+\mathcal{D}_{Style}$. This improves accuracy over $\mathcal{M}_{\mathcal{D}_R}$, but leads to a decrease of mAP compared to $\mathcal{M}_{\mathcal{D}_R+\mathcal{D}_M}$ (see Tab. 2). This result is surprising, as many factors believed to limit generalization are improved and the data is visually more realistic. A possible explanation is that the network learns to detect artifacts of the style transfer. Alternatively, the failure cases of the style transfer method might lead to "confusion" of the network.

In an ablation study, we test whether training on synthetic humans actually improves the mAP or if the improvement when training with $\mathcal{D}_M$ or $\mathcal{D}_{Style}$ is mostly caused by additional occlusion. For that purpose, we mask out all the loss that is generated by synthetic humans. Comparing $\mathcal{M}_{\mathcal{D}_R+\mathcal{D}_M+\text{masks}}$ with $\mathcal{M}_{\mathcal{D}_R+\mathcal{D}_M}$, it can be seen that masking out the loss generated by synthetic humans increases the accuracy. Therefore, the domain gap between synthetic and real humans limits the generalization, and the improvement of $\mathcal{M}_{\mathcal{D}_R+\mathcal{D}_M}$ over $\mathcal{M}_{\mathcal{D}_R}$ is due to more occlusion. Stylization in combination with masks leads to the best model (see Fig. 1 and **Supp. Mat.** for qualitative results), suggesting that a smaller gap between synthetic occluder and real parts of the image improves generalization.

5.3 Does Informed Sampling Improve Results?

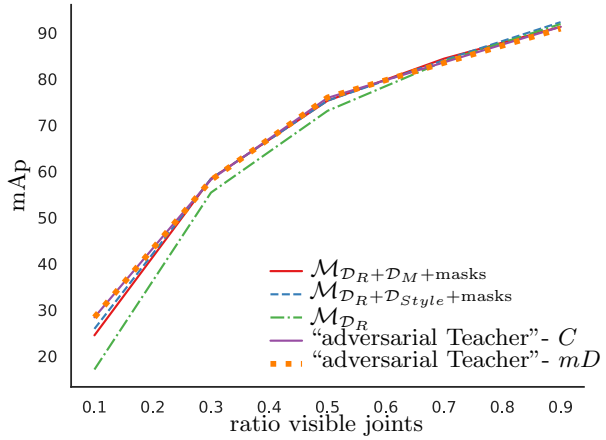
Finally, we test whether the teacher network can help to use the synthetic data more effectively. For that purpose we use the adversarial teacher with $\mathcal{D}_R+\mathcal{D}_S$. The results can be seen in Tab. 3. Grouping according to the camera pitch leads to an additional improvement of 0.5 mAP in comparison to $\mathcal{M}_{\mathcal{D}_R+\mathcal{D}_S}$. See Fig. 1 and **Supp. Mat.** for qualitative results. For usage of the teacher in combination with the mixed and stylized datasets we do not observe further improvements. Since we only consider real humans as samples some groups are very small. We assume that oversampling of these groups inhibits improvements.

As can be seen in Fig. 6, improvements for highly occluded people are strongest for models trained with the teacher. Clear differences to $\mathcal{M}_{\mathcal{D}_R+\mathcal{D}_M+\text{masks}}$

Table 3. Comparison of the $\mathcal{M}_{\mathcal{D}_R}$ and $\mathcal{M}_{\mathcal{D}_R+\mathcal{D}_S}$ baselines (copied from Tab. 1) to the student-teacher model using different groupings. Results on the validation set.

Model	Grouping	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	mAP
$\mathcal{M}_{\mathcal{D}_R}$		91.3	89.1	79.2	70.4	75.9	71.5	66.7	77.7
$\mathcal{M}_{\mathcal{D}_R+\mathcal{D}_S}$		91.1	89.2	80.5	71.0	75.2	73.6	**68.1**	78.4
"adversarial Teacher"	C	**91.7**	90.0	**80.9**	71.2	**77.1**	**73.6**	67.7	**78.9**
"adversarial Teacher"	mD	91.5	**90.4**	80.5	**72.2**	75.8	73.1	67.6	78.7

Fig. 6. Detection performance for varying ratio of visible joints (mAP). The validation data is grouped into 5 linearly spaced groups in range $[0, 1]$. The groups contain 14, 111, 289, 286, 180 persons respectively. The teacher methods are hard to distinguish because of similar values.



and $\mathcal{M}_{\mathcal{D}_R+\mathcal{D}_{Style}+\text{masks}}$ can only be seen for the highest occlusion level. For all but the lowest occlusion levels $\mathcal{M}_{\mathcal{D}_R}$ is outperformed by all other models. Thus, our methods of training improve accuracy for difficult high occlusion cases.

Sampling Probabilities. The teacher often assigns high probability to few groups early in training. In most cases the teacher converges to a uniform sampling strategy over the groups, as training progresses. This is not equivalent to random sampling, as samples are not uniformly distributed across groups. As a result the training data follows a uniform distribution for the respective image characteristic. A uniform distribution over an image characteristic like camera pitch is more extreme than the distribution in the real training data. Training on more extreme samples seems to improve the generalization. We find that improvements are largest for the camera pitch grouping. This might be due to the bias to small camera pitch in the real training data.

6 Conclusion

In summary, we created multiple synthetic datasets and analyze their effectiveness. We show that training with synthetic data improves multi-person pose estimation methods. We find that both our methods of generating synthetic datasets perform on par. Surprisingly, our approach for improving visual appearance of synthetic humans decreased the accuracy. More elaborate domain adaptation methods might provide better results. For example, GAN-based approaches ensure that no obvious artifacts can be used to distinguish real from synthetic. We find that improvements of the $\mathcal{M}_{\mathcal{D}_R+\mathcal{D}_M}$ and $\mathcal{M}_{\mathcal{D}_R+\mathcal{D}_{Style}}$ can be explained by more occlusion, as mAP increases when masking out the loss of synthetic humans. Here the stylization leads to further improvements, suggesting that visual quality of occluding objects is important. Finally, we find that training on the most difficult synthetic samples at each point of training improves the results. This suggest that, for large synthetic datasets, random sampling is not optimal and better strategies exist. More research in this direction is necessary to draw final conclusions. $\mathcal{M}_{\mathcal{D}_R+\mathcal{D}_{Style}+\text{masks}}$ outperforms the "adversarial Teacher" C model. We assume that the potential of the teacher is limited by the small amount and the quality of human textures. Given better textures we expect this approach to outperform $\mathcal{M}_{\mathcal{D}_R+\mathcal{D}_{Style}+\text{masks}}$.

Limitations. The teacher network is limited in multiple ways. First, the current implementation requires grouping of data. The size and spacing of groups might have a large influence on the training success and applicability of teacher networks. Furthermore, the grouping is based on one feature only. This is suboptimal, since difficulty of an image is determined by multiple factors. We plan to extend the teachers to handle multiple characteristics at once. A more elaborate formulation that does not require grouping might be superior. Last, the teacher can be applied to other tasks and networks, here we evaluate it only for multi-person pose estimation with the OpenPose network.

The style transfer occasionally produces artifacts in the stylized image. Especially the skin color of the synthetic humans might be unnatural. In rare cases, large parts of the background are included in the human mask. These failures in segmentation can lead to ghost-like synthetic humans (see **Supp. Mat.**).

Acknowledgement. S. Tang acknowledges funding by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) Projektnummer 276693517 SFB 1233.

Disclosure. MJB has received research gift funds from Intel, Nvidia, Adobe, Facebook, and Amazon. While MJB is a part-time employee of Amazon, his research was performed solely at, and funded solely by, MPI. MJB has financial interests in Amazon and Meshcapade GmbH.

References

1. Abdulla, W.: Mask r-cnn for object detection and instance segmentation on keras and tensorflow. https://github.com/matterport/Mask_RCNN (2017)
2. Alhaija, H.A., Mustikovela, S.K., Mescheder, L., Geiger, A., Rother, C.: Augmented reality meets computer vision: Efficient data generation for urban driving scenes. IJCV **126**(9), 961–972 (2018)
3. Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: 2d human pose estimation: New benchmark and state of the art analysis. In: CVPR (June 2014)
4. Bak, S., Carr, P., Lalonde, J.F.: Domain adaptation through synthesis for unsupervised person re-identification. In: ECCV (2018)
5. Barbosa, I.B., Cristani, M., Caputo, B., Rognhaugen, A., Theoharis, T.: Looking beyond appearances: Synthetic training data for deep cnns in re-identification. CVIU **167**, 50–62 (2018)
6. Bengio, Y., Louradour, J., Collobert, R., Weston, J.: Curriculum learning. In: ICML (2009)
7. Büchler, U., Brattoli, B., Ommer, B.: Improving spatiotemporal self-supervision by deep reinforcement learning. In: ECCV (2018)
8. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: CVPR (2017)
9. Chen, W., Wang, H., Li, Y., Su, H., Wang, Z., Tu, C., Lischinski, D., Cohen-Or, D., Chen, B.: Synthesizing training images for boosting human 3d pose estimation. In: 3DV. IEEE (2016)
10. Dundar, A., Liu, M.Y., Wang, T.C., Zedlewski, J., Kautz, J.: Domain stylization: A strong, simple baseline for synthetic to real image domain adaptation. arXiv preprint arXiv:1807.09384 (2018)
11. Dvornik, N., Mairal, J., Schmid, C.: On the importance of visual context for data augmentation in scene understanding. arXiv preprint arXiv:1809.02492 (2018)
12. Fabbri, M., Lanzi, F., Calderara, S., Palazzi, A., Vezzani, R., Cucchiara, R.: Learning to detect and track visible and occluded body joints in a virtual world. In: ECCV (2018)
13. Fan, Y., Tian, F., Qin, T., Bian, J., Liu, T.Y.: Learning what data to learn. arXiv preprint arXiv:1702.08635 (2017)
14. Fang, H., Xie, S., Tai, Y.W., Lu, C.: Rmpe: Regional multi-person pose estimation. In: ICCV (2017)
15. Fieraru, M., Khoreva, A., Pishchulin, L., Schiele, B.: Learning to refine human pose estimation. In: CVPR Workshops (2018)
16. Ghezelghieh, M.F., Kasturi, R., Sarkar, S.: Learning camera viewpoint using cnn to improve 3d body pose estimation. In: 3DV (2016)
17. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: ICCV (2017)
18. Hinterstoisser, S., Lepetit, V., Wohlhart, P., Konolige, K.: On pre-trained image features and synthetic images for deep learning. In: ECCV (2018)
19. Insafutdinov, E., Pishchulin, L., Andres, B., Andriluka, M., Schiele, B.: Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In: European Conference on Computer Vision. pp. 34–50. Springer (2016)
20. Jiang, L., Zhou, Z., Leung, T., Li, L.J., Fei-Fei, L.: Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. arXiv preprint arXiv:1712.05055 (2017)
21. Katharopoulos, A., Fleuret, F.: Biased importance sampling for deep neural network training. arXiv preprint arXiv:1706.00043 (2017)

22. Kim, T.H., Choi, J.: Screenernet: Learning self-paced curriculum for deep neural networks. arXiv preprint arXiv:1801.00904 (2018)
23. Kocabas, M., Karagoz, S., Akbas, E.: Multiposenet: Fast multi-person pose estimation using pose residual network. In: ECCV (2018)
24. Kumar, M.P., Packer, B., Koller, D.: Self-paced learning for latent variable models. In: NIPS (2010)
25. Li, Y., Liu, M.Y., Li, X., Yang, M.H., Kautz, J.: A closed-form solution to photorealistic image stylization. In: ECCV (2018)
26. Loper, M., Mahmood, N., Black, M.J.: Mosh: Motion and shape capture from sparse markers. TOG **33**(6), 220 (2014)
27. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. TOG **34**(6), 248 (2015)
28. Luo, Y., Xu, Z., Liu, P., Du, Y., Guo, J.M.: Multi-person pose estimation via multi-layer fractal network and joints kinship pattern. TIP **28**(1), 142–155 (2019)
29. Marin, J., Vázquez, D., Gerónimo, D., López, A.M.: Learning appearance in virtual scenarios for pedestrian detection. In: CVPR (2010)
30. Müller, M., Casser, V., Lahoud, J., Smith, N., Ghanem, B.: Sim4cv: A photorealistic simulator for computer vision applications. IJCV pp. 1–18 (2018)
31. Murez, Z., Kolouri, S., Kriegman, D., Ramamoorthi, R., Kim, K.: Image to image translation for domain adaptation. In: CVPR (2018)
32. Newell, A., Huang, Z., Deng, J.: Associative embedding: End-to-end learning for joint detection and grouping. In: NIPS (2017)
33. Nie, X., Feng, J., Xing, J., Yan, S.: Pose partition networks for multi-person pose estimation. In: ECCV (2018)
34. Peng, X., Tang, Z., Yang, F., Feris, R.S., Metaxas, D.: Jointly optimize data augmentation and network training: Adversarial data augmentation in human pose estimation. In: CVPR (2018)
35. Pishchulin, L., Jain, A., Wojek, C., Andriluka, M., Thormählen, T., Schiele, B.: Learning people detection models from few training samples. In: CVPR (2011)
36. Ranjan, A., Romero, J., Black, M.J.: Learning human optical flow. In: BMVC (2018)
37. Rogez, G., Schmid, C.: Image-based synthesis for deep 3d human pose estimation. IJCV **126**(9), 993–1008 (2018)
38. Rogez, G., Weinzaepfel, P., Schmid, C.: Lcr-net++: Multi-person 2d and 3d pose detection in natural images. TPAMI (2019)
39. Romero, J., Tzionas, D., Black, M.J.: Embodied hands: Modeling and capturing hands and bodies together. ACM TOG, (Proc. SIGGRAPH Asia) **36**(6) (Nov 2017)
40. Sárándi, I., Linder, T., Arras, K.O., Leibe, B.: How robust is 3d human pose estimation to occlusion? arXiv preprint arXiv:1808.09316 (2018)
41. Shrivastava, A., Gupta, A., Girshick, R.: Training region-based object detectors with online hard example mining. In: CVPR (2016)
42. Shrivastava, A., Pfister, T., Tuzel, O., Susskind, J., Wang, W., Webb, R.: Learning from simulated and unsupervised images through adversarial training. In: CVPR (2017)
43. Tripathi, S., Chandra, S., Agrawal, A., Tyagi, A., Rehg, J.M., Chari, V.: Learning to generate synthetic data via compositing. arXiv preprint arXiv:1904.05475 (2019)
44. Varol, G., Romero, J., Martin, X., Mahmood, N., Black, M.J., Laptev, I., Schmid, C.: Learning from synthetic humans. In: CVPR (2017)
45. Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.: Sun database: Large-scale scene recognition from abbey to zoo. In: CVPR (2010)