GENERALIZED POSE EMBEDDINGS FOR TRAINING IN-THE-WILD VIA ANALYSIS-BY-SYNTHESIS

Dominik Borer
ETH Zurich
Disney Research | Studios

Jakob Buhmann
Disney Research | Studios

Martin Guay
Disney Research | Studios

October 15, 2022

ABSTRACT

Modern pose estimation models are trained on large, manually-labelled datasets which are costly and may not cover the full extent of human poses and appearances in the real world. With advances in neural rendering, analysis-by-synthesis and the ability to not only predict, but also render the pose, is becoming an appealing framework, which could alleviate the need for large scale manual labelling efforts. While recent work have shown the feasibility of this approach, the predictions admit many flips due to a simplistic intermediate skeleton representation, resulting in low precision and inhibiting the acquisition of any downstream knowledge such as three-dimensional positioning. We solve this problem with a more expressive intermediate skeleton representation capable of capturing the semantics of the pose (left and right), which significantly reduces flips. To successfully train this new representation, we extend the analysis-by-synthesis framework with a training protocol based on synthetic data. We show that our representation results in less flips and more accurate predictions. Our approach outperforms previous models trained with analysis-by-synthesis on standard benchmarks.

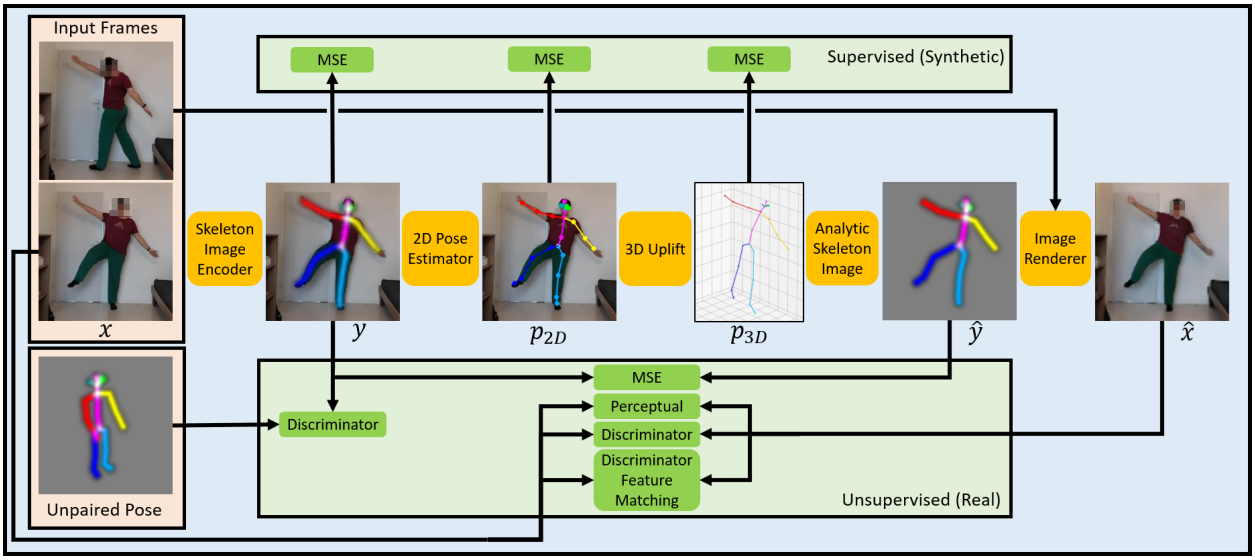

Figure 1: Overview of the model and training procedure. There are 5 main components (orange). First, the input image x is mapped to a skeleton image representation y. From this the 2D pose p_{2D} is estimated, which is then uplifted to 3D joint positions and orientations p_{3D}. The reprojected coordinates are then used to analytically create a skeleton image $\hat{y}$, from which the input image is reconstructed $\hat{x}$. To train the model we use a mixture of synthetic and real data to optimize several objectives (green).

1 Introduction

Building datasets for pose estimation of humans and animals is costly and remains challenging in terms of capturing the full diversity in pose and appearance of the real world. Some works seek to utilize synthetic humans but remain at a gap from real world imagery. This inhibits the ability of labelled datasets—both synthetic and real—to generalize and motivates the need for a class of methods that can learn from un-labelled video data in-the-wild.

Analysis-by-synthesis is the idea of including a depiction model in order to define a loss or cost back to the input image—enabling a training signal directly from unlabelled images. Such an approach, within a deep learning human pose estimation setting has been demonstrated in Jakab et al. [2020], where they use an intermediate pictorial skeleton representation of the pose. While their representation of a pose can be trained in an unsupervised way, it consists of only a single channel ($\mathbb{R}^{1\times W\times H}$) and does not represent the semantics of body parts—leading to significant flips between left and right, front and back in both the pose predictions and renderings of the person, as can be seen in Figure 3.

We solve this problem with a new multi-channel pose representation. However, we found this higher capacity model sensitive to divergence during training. To address this problem, we introduce synthetic humans to pre-train our pose and rendering model, providing a better conditioning for further fine-tuning using analysis-by-synthesis in the wild. Another way to look at our approach is to consider analysis-by-synthesis a tool for bridging the reality gap of models trained with synthetic data.

On the standard benchmark Human3.6M Ionescu et al. [2014], we outperform the prior work of Jakab Jakab et al. [2020] with an MSE of 10.39, compared to the MSE of 14.46 for the baseline. We also gathered specialized data of a target subject and measured whether fine-tuning on this data would improve accuracy at run-time. On the Human3.6M benchmark, this refinement step further reduces the MSE to 6.62, outperforming the baseline by a large margin. Additionally, we investigated our framework further to answer whether a 3D pose representation could be trained end-to-end. We also show that our approach can generalize to other skeletal structures such as animals, where we qualitatively improve the accuracy of the 3D pose predictions compared to the work of Borer Borer et al. [2021a].

2 Related Work

Our work sits at the intersection of supervised and un-supervised pose estimation, where neural rendering enables learning from un-labelled images, while supervision is provided by synthetic data, both to initialize our model, and to provide a noise-free pose prior.

Fully supervised methods leverage large-scale datasets with 2D and/or 3D annotations such as MS COCO Lin et al. [2014], Human3.6M Ionescu et al. [2014], CMU CMU [2001], and MPII Andriluka et al. [2014]. Such approaches usually do not need an additional pose prior, as it is already captured in the data. Approaches for 2D pose estimation use CNNs to regress keypoint confidence maps Wei et al. [2016], Newell et al. [2016], which can be refined iteratively Cao et al. [2018], or directly regress keypoint coordinates Toshev and Szegedy [2014]. For 3D pose estimation it is common to train a separate model to go from 2D keypoints to 3D joint locations and orientations Martinez et al. [2017], but others also directly regress to 3D poses from images Borer et al. [2021a]. Because hand-labelling datasets is very expensive and motion capture datasets lack variation in the images, other works train from synthetically generated data Varol et al. [2017]. Even though not as accurate as training on real data, this has proven to be practical for exposing DNNs to 3D information, as well as to domains such as animals, where annotated data is not available or very hard to acquire Zuffi et al. [2017, 2019], Borer et al. [2021a]. In this work, we also rely on synthetic data, but combine it with unlabelled real data to address the reality gap.

Weak Supervision Methods with weak supervision assume that only some annotations are available. In Kanazawa et al. [2018], Pavlakos et al. [2018], Habermann et al. [2020] a dense 3D human mesh is predicted from sparse 2D keypoint annotations, with a loss on the reprojected keypoints. To ensure a plausible 3D shape, the SMPL parametric human mesh model Loper et al. [2015] is used as a prior. This is combined with motion capture data for adversarial learning Kanazawa et al. [2018], Yang et al. [2018], a loss on the rendered silhouette Pavlakos et al. [2018], or multi-view consistency Habermann et al. [2020]. Similarly, we also use motion capture data for an empirical pose prior, but we do not need a fully fledged 3D prior such as provided by the SMPL model.

Un-supervised Pose Estimation On the other hand, unsupervised methods do not need any annotations. Works such as Kanazawa et al. [2016] learn to match pairs of images of an object, but do not learn geometric information like the pose. Landmarks as an explicit structural representation can be learned through analysis-by-synthesis, by reconstructing the input image conditioned on the landmarks Thewlis et al. [2017], Jakab et al. [2018], Zhang et al. [2018], Lorenz et al. [2019]. But these landmarks are not constrained to any semantics. Hence, additional paired

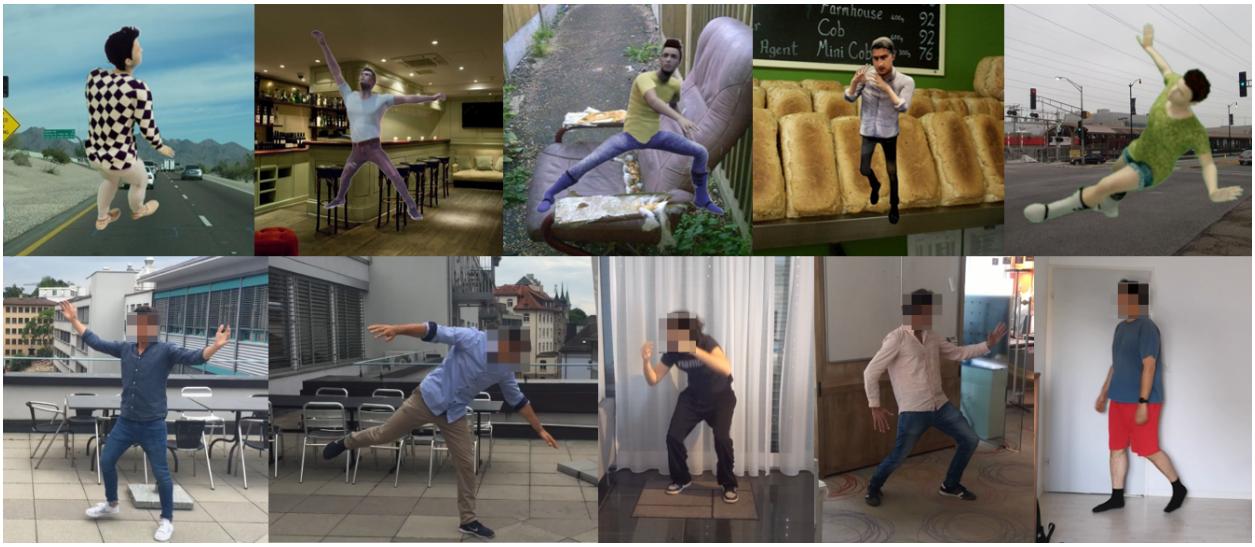

Figure 2: Samples from the training dataset. Top: Synthetically generated data. Bottom: Unlabelled, real, in-the-wild videos. The synthetic data contains a lot of variation in pose, appearance and background and the real data covers a variety of different people performing various motions.

supervision is required to map the landmarks to the desired semantics. Other approaches tackle this lack of semantics by fitting a template model to the landmarks Kundu et al. [2020], Schmidtke et al. [2021]. Directly using a pose prior inside the network, forces the landmarks to follow the desired pose representation without the need for a template model Jakab et al. [2020]. While this directly learns a pose structure, it remains limited to a single channel and admits significant left/right ambiguities. We solve this problem in this work with a richer pose model that is directly learned and shows an increase in performance.

Another line of work in unsupervised pose estimation relies on multi-view consistency Rhodin et al. [2018], Chen et al. [2019]. Given the predicted pose from one view, the image in a different view is reconstructed Rhodin et al. [2018]. Or for monocular data the predicted 3D pose is projected to a different camera and uplifted again to compare to the original pose Chen et al. [2019]. These training schemes are orthogonal to our work and could easily be integrated into our framework.

Adversarial Learning - Neural Rendering For the approach of analysis-by-synthesis to work, the original input image has to be reconstructed, which creates the training signal. This unsupervised training falls into the domain of neural rendering and generative imagery. To achieve this, adversarial learning Goodfellow et al. [2014] is useful to train a deep generative network to translate images from one domain to another using paired Isola et al. [2017] or unpaired Zhu et al. [2017] samples, and is further improved to high resolution images in Wang et al. [2018a], Park et al. [2019a]. Similar ideas are also applied to articulated characters by translating an image of a skeleton Chan et al. [2019], Aberman et al. [2019] or a segmentation map Sarkar et al. [2021] to an image of a person or to directly map a 3D pose vector to a high resolution image Borer et al. [2021b]. In our case, we use image-to-image translation networks to predict the skeleton image and to reconstruct the input image.

3 Dataset

To address the data annotation issue, we use a mixture of synthetic data as well as unlabelled real-world data. This way we do not rely on manual labelling, and use the synthetic data to pre-train our model, while images-in-the-wild are used to fine tune our network to real world scenarios.

To generate the synthetic data, we use a setup similar to Borer et al. [2021a]. We use domain randomization Tobin et al. [2017] to create variations in appearance, shape and pose, as can be seen in the top row of Figure 2. As shown in Borer et al. [2021a], a model trained on such data has the ability to detect poses in the real-world, but it remains at a gap from models trained on real-world data, which we provide in this work through analysis-by-synthesis.

Our unlabelled real-world data consists of videos in-the-wild comprising different people, performing various actions in different environments, as can be seen in the bottom row of Figure 2. While we do not strictly require video data, we

Figure 3: The single-channel skeleton image representation Jakab et al. [2020] suffers from ambiguities and fails to capture the body part semantics, causing flips in the predicted pose (e.g. red and blue should be the right arm and leg).

need at least two frames of the same person in the same environment. We employ a pose prior that is trained on an in-house 3D motion capture dataset similar to Human3.6M Ionescu et al. [2014]. Using the synthetic data generation pipeline, we extend this pose data with additional surface keypoints, i.e. facial keypoints, to further help with the ambiguity issue.

4 Method

Our goal is to train a neural network that estimates the 2D and 3D pose of a person, given a single monocular image. Following the analysis-by-synthesis paradigm, we first extract the pose of the person from the input image and then reconstruct the image using this pose data as input. With a naive auto-encoder formulation, given enough freedom, the model can simply output a copy of the input image, without learning anything useful. To prevent that, we use a dual representation of the pose as a vector of keypoint coordinates and a skeleton image, similar to the one introduced in Jakab et al. [2020]. This dual representation creates a tight bottleneck that disentangles pose from appearance, and allows the use of image-to-image translation networks to reconstruct the input image. Separating pose from appearance allows the model to specialize on a single task, improving the performance. Since the pose alone lacks appearance information, the image reconstruction is ill posed, and hence, an additional but different image of the same subject is provided as the appearance code. Furthermore, to ensure the predicted poses are plausible, the distribution of the pose space is constrained using a pose prior, learned from the unpaired pose data through adversarial training.

Overall, our model consists of five components, as depicted in Figure 1. First, the *Skeleton Image Encoder* maps the input image to the skeleton image representation. The *2D Pose Estimator* and *3D Uplift* then predict keypoint coordinates and 3D joint positions as well as orientations. After reprojecting the joint positions, we create the *Analytic Skeleton Image*, from which the *Image Renderer* then reconstructs the input image.

4.1 Multi-Channel Skeleton Image

The skeleton image y, introduced in Jakab et al. [2020], is a pictorial representation of the pose, where each pixel's intensity is computed from the distance to the closest limb:

$$y = \exp\left(-\gamma \min_{\substack{(i,j)\in E \\ t\in[0,1]}} ||u - ((1-t)\cdot p_i + t\cdot p_j)||^2\right), \tag{1}$$

where E is the set of connected keypoint pairs (i, j), p a keypoint position, u an image pixel coordinate, and γ a scaling factor (we used $\gamma = 250$ for our experiments). To solve the left/right ambiguity of the single-channel skeleton image

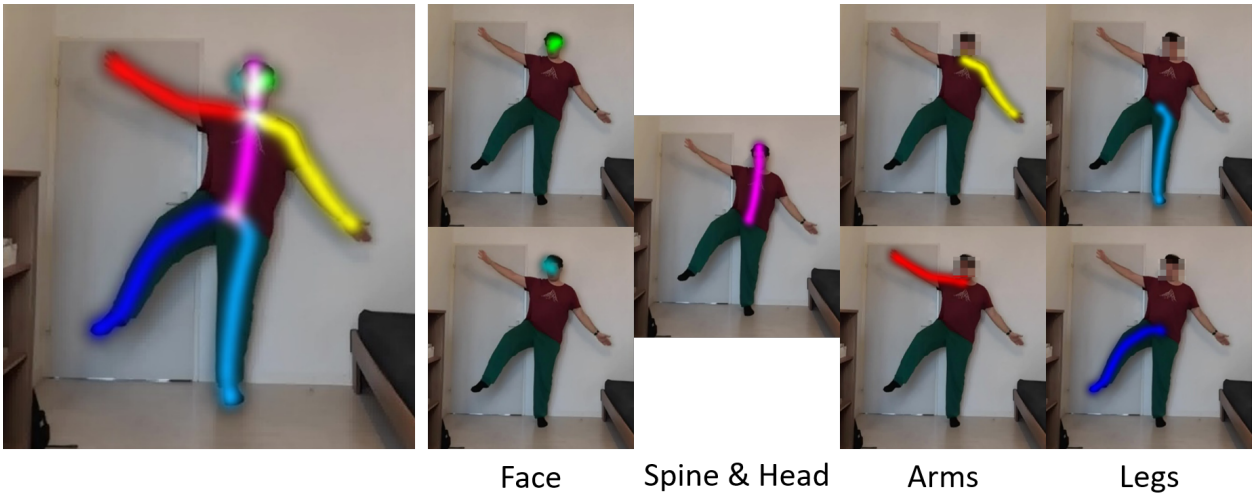

Figure 4: Our multi-channel skeleton image representation. Each channel (visualized with different colors) represents a semantically meaningful set of joints.

shown in Figure 3, we extend this representation to a multi-channel image $y \in \mathbb{R}^{C \times W \times H}$, where C is the number of channels. Each channel represents a semantically meaningful set of joints. While the exact separation and number of channels may vary, we found empirically that a separation into left and right sides, as well as individual limbs as depicted in Figure 4 yielded best performance. This representation is powerful enough to resolve the ambiguity issues as shown in Figure 5, without requiring any changes to the model architecture/capacity nor drastically increase the memory or computation requirements.

Figure 5: Predicted poses when using our multi-channel skeleton image. The predictions are accurate for a wide range of poses and do not suffer from left/right flips.

While our new pose representation should reduce left and right ambiguities, the unsupervised training used in Jakab et al. [2020] lacks conditioning to train successfully. The result is a model capable of generating the input image, without learning keypoints that match the pose in the image. To solve this problem, we pre-train our model with labelled data that was computer-generated, providing the initial conditioning required for further fine tuning.

4.2 Uplifting to 3D

Prior works using analysis-by-synthesis for pose estimation, only predict 2D poses. And supervised approaches for 3D pose estimation commonly train a separate 2D-to-3D uplift model due to lack of images with 3D annotations. Since such a model is trained from ground truth 2D keypoints, it does not generalize to the noise of a 2D pose estimator, and modelling that noise during the training is hard. Consequently, the 3D pose is not as accurate and does not overlap well with the image. To improve this, we include an uplift module Martinez et al. [2017] in the middle of our model (see Figure 1) and train it end-to-end. Specifically, we uplift the predicted 2D keypoints and then reproject the 3D joint positions before creating the analytic skeleton image. For the reprojection, we use a perspective camera. Since we do not know the intrinsic camera parameters, we fix them to plausible defaults (i.e. field of view of 62°). The image reconstruction loss then pushes the model towards 3D poses that overlap with the image, as can be seen in Figure 6.

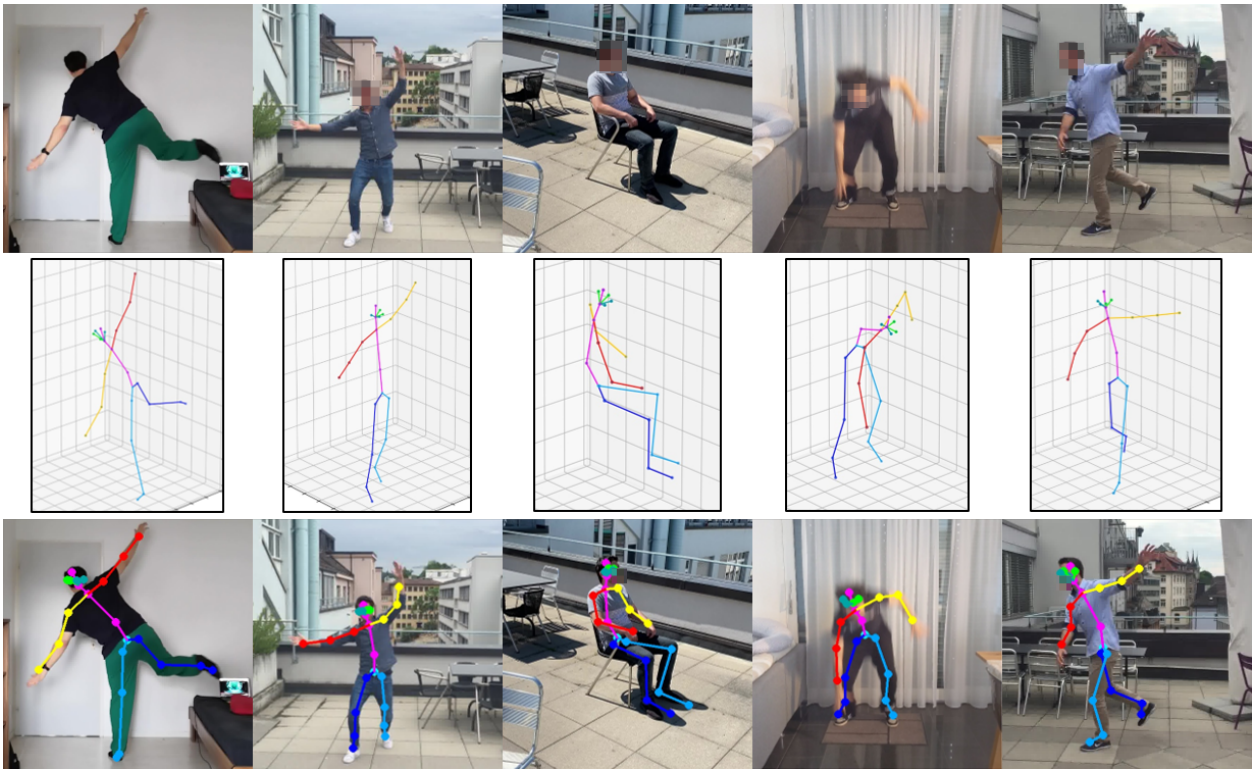

Figure 6: Uplifted 3D poses. Due to the end-to-end training, the poses are expressive and the reprojection (last row) overlaps with the person in the input image.

5 Training

We use a mixture of synthetic and unlabelled real data in two steps. First, we use the synthetic data for pretraining to bootstrap the model, then train in an unsupervised way on real data, and lastly (as well as optionally) improve the performance for a target video with an unsupervised instance-specific refinement step.

Pretraining on Synthetic Data To bootstrap the model, we pretrain it in a supervised way. With the synthetic data, we can pretrain the individual components independently, which converges faster than training end-to-end. The pretrained components might not work well together, but we found that it is not necessary to refine the model end-to-end on synthetic data. Even though the pretrained model has difficulties to generalize to real data, despite domain randomization, we found that it suffices for bootstrapping the model for the unsupervised training.

Unsupervised Training on Real Data With the unlabelled real data, we then continue with the end-to-end training in an unsupervised way. While the image data changes, we reuse the same pose data from before for the pose prior. Besides optimizing several objectives described in Section 5.1, we continue with synthetic supervision, but with a reduced weight. This helps to regularize and better condition the training, without constraining the model to the synthetic domain. The additional supervision prevents the model from diverging and learning keypoints that do not match the pose, while the input image is successfully rerendered. The trained model then generalizes well to similar real data.

Instance-Specific Refinement Given a completely new, unseen video, depending on the similarity to the training data, the model might still struggle e.g. for a different person in a different environment. To address that we use an instance-specific refinement step. To refine the model for a particular person we first collect a few videos of this person, performing various motions. Starting from the trained model, we replace the unlabelled, real training data with the videos of the target person and continue training for a short time. The refined model then generalizes better to videos of the target person. Alternatively, we can refine on a single target video, by using just this video as the unlabelled training data, which considerably improves the performance, as shown in Figure 8. In this case the refinement can also be seen as a post-process optimization step.

5.1 Training Objectives

For the unsupervised training to succeed, we optimize several objectives, as shown in Figure 1. This includes losses on the reconstructed image (render loss), losses on the predicted pose (pose prior) as well as losses on synthetic data.

Render Loss To train our model, we use a dataset of N images $\{x_i\}_{i=1}^N$ to optimize the reconstruction loss. Instead of directly comparing pixels we use the perceptual loss Dosovitskiy and Brox [2016], which compares features extracted from different layers of a pretrained feature extractor Γ, such as VGG Simonyan and Zisserman [2015]:

$$L_{perc\_img} = \frac{1}{N}\sum_{i=1}^{N} ||\Gamma_l(x_i) - \Gamma_l(\hat{x}_i)||_2^2 \,, \tag{2}$$

where $\hat{x}_i$ is the reconstructed image and Γ_l the extracted features at layer l. The losses for the different layers are averaged. Additionally, we use adversarial training with a multi-scale discriminator D Wang et al. [2018b] to capture features at different scales (1, $\frac{1}{2}$, $\frac{1}{4}$). For the discriminator loss we employ a least square loss Mao et al. [2017]:

$$L_{disc\_img} = \sum D(x_{real})^2 + \sum (1 - D(x_{fake}))^2. \tag{3}$$

Similar to the perceptual loss we also use a discriminator feature matching loss Park et al. [2019b], where the intermediate features of the discriminator are compared:

$$L_{disc\_img\_FM} = \frac{1}{N}\sum_{i=1}^{N} |D_l(x_i) - D_l(\hat{x}_i)| \,, \tag{4}$$

where D_l are the features at layer l. Both discriminator losses are averaged over the different scales.

Pose Prior Besides the unlabelled images, we use a dataset of M unpaired poses from which we create skeleton images $\{\bar{y}\}_{i=1}^M$ with Equation 1 to build a pose prior to encourage the model to predict plausible poses y. Similar to the render loss, we use a multi-scale discriminator D_{sk} for the skeleton images y, with a least squares discriminator loss:

$$L_{disc\_sk} = \sum D_{sk}(y_{real})^2 + \sum (1 - D_{sk}(y_{fake}))^2. \tag{5}$$

Additionally, to ensure the estimated 2D and 3D pose follows our desired structure, we have an L2 reconstruction loss L_{rec_sk} between the predicted skeleton image y and the analytically reconstructed skeleton image $\hat{y}$ from the keypoint 2D coordinates as well as the reprojected 3D positions (for simplicity Figure 1 shows only the skeleton image created from the reprojected positions).

Synthetic Supervision During pretraining of the individual modules, we use L2 losses on the predicted skeleton image y, the 2D coordinates p_{2D}, and the 3D positions and orientations p_{3D}. For the reconstructed image, we use the same discriminator losses as explained above. During the training on real data, we keep only the L2 losses shown in Figure 1. The discriminator loss would push the renderer towards synthetic images, which only harms the training.

Overall Learning Objective Combining all losses and balancing their contributions yields the overall objective (the weights can be found in the supplementary material). Similar to any adversarial formulation the loss on the unlablled data is maximized the two discriminators and minimized the other components.

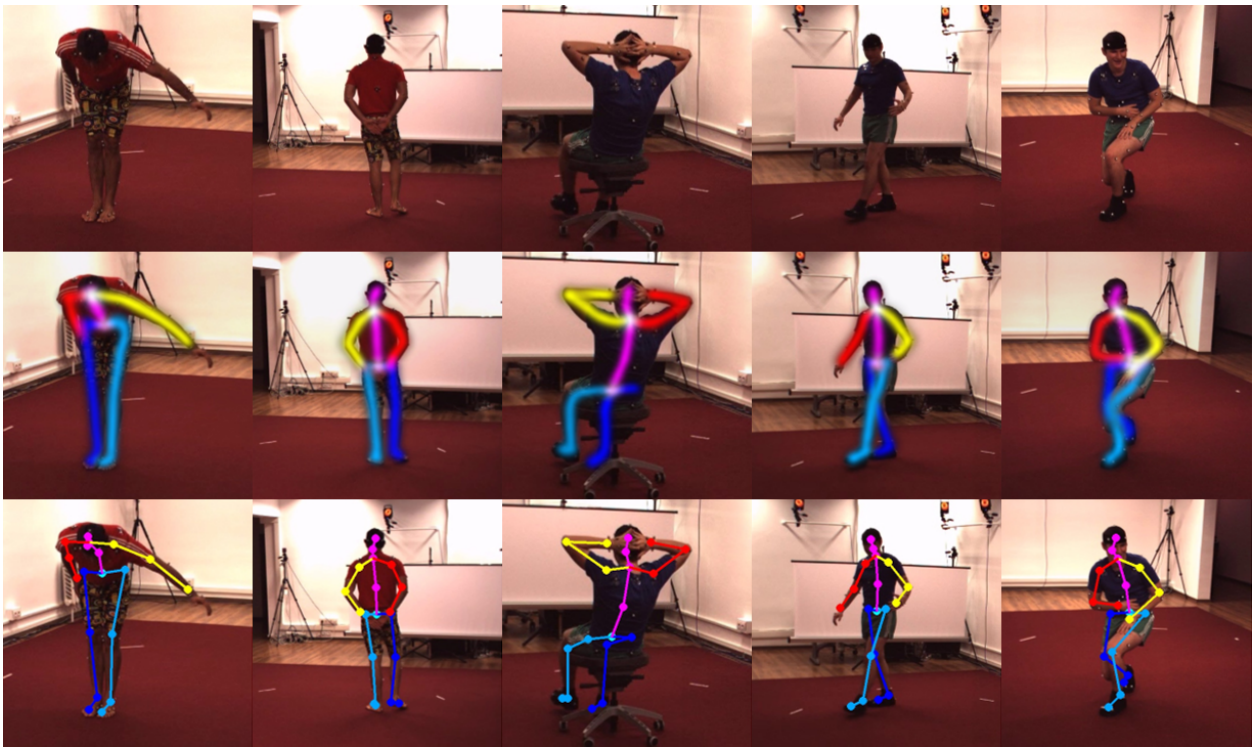

Figure 7: Results on the Human3.6M test set. The predictions are accurate and have no ambiguity issues.

5.2 Data Augmentation

The prior work of Jakab et al. [2020] always assumes a tight crop and thus the person will always roughly be at the same scale, which significantly simplifies the problem. But for in-the-wild videos, where no crop is available, such a system will fail. Similarly, if the training data contains only a few different proportions (e.g. as in the Human3.6M dataset), the model will have a hard time generalizing to new proportions. This is especially problematic for the 2D-to-3D uplifting, where the predicted 3D pose will be wrongly deformed along the depth axis in order for the reprojection to match.

To generalize better to those cases, we rely on data augmentation. To address the issue with the scale, we randomize the size and offset of the crop. And to address the issue with the proportions, we randomize the limb lengths of the pose data in 3D space and reproject them to get the corresponding 2D coordinates. This way the model can generalize better to different scales and proportions. This also resolves proportion mismatches between the pose prior and the real training data. If we do not randomize the proportions, the model will for example consistently predict a smaller skeleton, but with the right pose. The image renderer then compensates for this scale difference to minimize the render loss. Since there is no overlap for such predictions, they are not usable.

6 Evaluation

To evaluate the performance of our model, we compare our multi-channel representation to the single-channel representation of Jakab et al. [2020] on the Human3.6M dataset. Further we show qualitative results for in-the-wild videos, as well as other domains such as animals.

6.1 Human3.6M

The Human3.6M dataset Catalin Ionescu [2011], Ionescu et al. [2014] contains 3.6 million 2D and 3D human pose annotations for 15 different activities, recorded from 4 different viewpoints with static backgrounds. Following the standard protocol, we use subjects 1, 5, 6, 7, and 8 for training and subjects 9 and 11 for evaluation Villegas et al. [2017]. To compare to prior work, we follow the same scheme as Jakab et al. [2020] and use a disjoint set of the dataset for the pose prior. Because this data has no facial keypoints, we use only a 5-channel skeleton image. While a pose prior from from our in-house motion capture data, containing facial keypoints, also works (as shown with the in-the-wild results),

(ignore flip)	Jakab et al. [2020]	Ours			Ours Refined			Ours No-Synth (5-ch)	
	1-ch	1-ch	3-ch	5-ch	1-ch	3-ch	5-ch	Unrefined	Refined
direct	11.26	6.38	4.73	4.83	6.19	4.26	**2.98**	35.51	33.33
discuss	13.72	9.52	6.81	7.18	8.35	7.21	**4.93**	49.06	43.10
eat	12.02	8.12	6.23	6.21	7.59	6.20	**5.97**	55.41	50.45
greet	11.85	7.91	6.13	6.11	6.61	5.58	**3.66**	47.67	43.21
phone	14.42	17.94	10.70	10.90	16.39	10.97	**8.17**	75.52	72.62
pose	10.39	5.48	5.13	6.53	5.53	4.59	**2.67**	37.29	33.87
purchases	12.90	20.81	10.60	10.44	15.59	8.47	**4.77**	66.81	51.62
sit	17.01	32.87	22.30	25.24	28.01	27.09	**13.86**	146.53	139.44
sit down	25.71	78.13	37.18	37.52	73.32	29.54	**22.65**	191.83	135.04
smoke	14.35	19.27	10.29	9.15	17.19	9.98	**7.19**	77.54	71.20
take photo	18.67	16.58	9.07	8.44	15.73	8.97	**7.72**	67.13	57.96
wait	11.40	12.00	7.35	6.53	8.33	5.96	**4.12**	44.18	39.36
walk	11.85	6.13	4.52	5.32	5.90	4.18	**3.09**	47.77	46.48
walk dog	19.42	11.66	6.81	6.80	10.22	6.56	**4.55**	52.00	43.38
walk together	11.90	7.31	4.82	4.65	6.18	3.84	**2.91**	49.45	44.80
all	14.46	17.34	10.18	10.39	15.41	9.56	**6.62**	69.58	60.39

Table 1: Comparison of the MSE in pixel to prior work on the Human3.6M test set. We compare the performance of different configurations (*1-*, *2-* and *3-channels*), the improvements through the refinement, as well as the importance of the synthetic supervision (*No-Synth*). After training on real data (*Ours*) we achieve better or similar scores and after refining (*Ours Refined*) we outperform the baseline by a large margin. Configurations with more channels show to be superior over a single channel. Without synthetic supervision (*Ours No-Synth*) the training fails to converge to a good solution, leading to poor accuracy. Note that as in Jakab et al. [2020] this metric ignores flips by taking the smaller error of the current and the flipped predictions. In Table 2 we compare the results where flips are considered.

the differences in the skeleton structure prevent a proper comparison to prior work (e.g. offsets in keypoint locations lead to offsets in the error metric). Figure 7 shows qualitatively the accuracy of our 5-channel model on the test set. To validate the effectiveness of our approach we compare different skeleton image representations (1, 3 and 5 channels), before and after refinement, and the importance of the synthetic supervision. The results summarized in Table 1 and Table 2 show that our approach outperforms the single-channel baseline Jakab et al. [2020] and significantly reduces flips. After the unsupervised training on real data we achieve similar or better scores for almost all activities and after refining we outperform the baseline by a large margin.

6.2 In-the-Wild Videos

Whereas for annotated datasets we can crop around the person for optimal performance, for in-the-wild footage this is not available. Since we used in-the-wild videos for our real dataset, combined with data augmentation, our model is more robust to such cases. As can be seen in Figure 8 and our accompanying video, the estimated poses for people used during training are very accurate. While for a completely different person the model might have some difficulties, it can quickly adapt with the instance-specific refinement.

6.3 Animals

Our framework is not specific to humans, but can also work for other skeleton structures, like dogs or lions. For wild animals it is even more inherent that capturing the motion with traditional motion capture setups is difficult. Other works such as Zuffi et al. [2019], Borer et al. [2021a] only used synthetic data to generalize to wild animals, but this remains challenging due to the reality gap and in many cases the estimated 3D poses do not overlap well with the image. With our approach on the other hand we can leverage real footage of wild animals and the estimated 3D poses are more accurate and have a better overlap due to the end-to-end training, as shown in Figure 9 and our accompanying video. Furthermore, with the instance-specific refinement we can easily adapt to other breeds e.g. a model trained only on lions can be refined on a video of a dog.

7 Limitations and Future Work

One of the limitations of our model is the resolution of the rendering module (128×128 pixels). Increasing this resolution and quality of the rendered image, while challenging, could help further increase the effectiveness of the

(consider flip)	Jakab et al. [2020]	Ours			Ours Refined			Ours No-Synth (5-ch)	
	1-ch	1-ch	3-ch	5-ch	1-ch	3-ch	5-ch	Unrefined	Refined
direct	11.26	30.69	5.83	8.53	24.41	4.94	**3.44**	73.74	44.98
discuss	13.72	41.05	9.16	12.25	32.52	8.95	**5.98**	90.11	55.64
eat	12.02	52.76	7.72	10.43	43.19	6.89	**6.36**	80.89	60.90
greet	11.85	34.81	7.82	11.47	27.69	6.34	**4.56**	76.33	52.36
phone	14.42	50.77	12.75	16.97	41.75	11.98	**9.33**	108.59	93.08
pose	10.39	40.67	7.36	16.82	31.92	5.16	**2.80**	90.09	44.07
purchases	12.90	89.88	12.71	17.28	72.54	8.72	**4.84**	99.13	63.61
sit	17.01	93.59	34.19	35.72	82.45	37.02	**16.45**	193.50	191.91
sit down	**25.71**	161.64	58.27	64.30	160.72	40.22	28.66	305.54	171.18
smoke	14.35	60.68	15.02	14.61	50.48	11.98	**8.66**	120.18	97.89
take photo	18.67	70.82	12.85	13.09	62.46	10.23	**9.41**	108.12	80.05
wait	11.40	47.06	10.64	15.43	39.95	6.53	**4.45**	85.68	52.57
walk	11.85	35.43	7.00	11.08	28.46	4.45	**3.26**	81.67	60.82
walk dog	19.42	40.08	9.42	10.71	36.52	7.72	**5.07**	87.84	65.11
walk together	11.90	35.74	7.31	10.65	29.28	4.07	**2.97**	93.87	54.46
all	14.46	59.04	14.54	17.96	50.96	11.68	**7.75**	113.02	79.24

Table 2: Comparison of the MSE in pixel on the Human3.6M test set, where the metric does not ignore flips, in contrast to Jakab et al. [2020]. Compared to Table 1 we observe a much higher error for the single-channel configuration (*1-ch*), which shows that the single-channel skeleton representation is not expressive enough and suffers from many flips. For the multi-channel skeleton configurations (*3-ch*, *5-ch*) the difference is much smaller, especially after the refinement, which shows the effectiveness of the multi-channel representation.

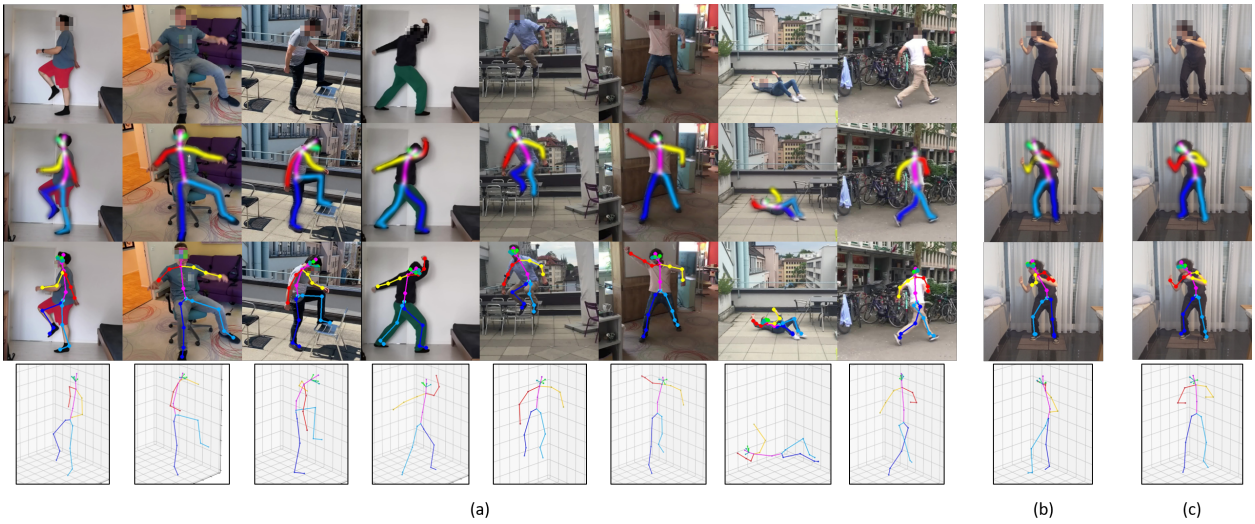

Figure 8: Predicted 2D and 3D poses for in-the-wild videos. Our approach can accurately estimate the pose of the person. For people used during training (a), it works very well. For a new, completely different person in a new environment (b), the model may have difficulties, but can quickly adapt through the unsupervised instance-specific refinement (c).

analysis-by-synthesis approach. Additionally, it would also be interesting to explore other image modalities such as UV, normal and depth maps and integrate them into the pipeline to help with the pose estimation as well as with the rendering.

Although we extended the pose representation to multiple channels, we are challenged to extend further, say to one channel per bone. We believe future work with a tailored architecture and training curriculum could make this feasible.

Besides the use case of monocular pose estimation, it would be interesting to explore the effectiveness of analysis-by-synthesis in a multi-camera scenario, where we could make use of multi-view consistency. A multi-camera setup is often available in lab scenarios, for example to study the behaviour of animals as in Bala et al. [2020]. But currently employed techniques still rely on large, manually annotated datasets of the animal to study, which is expensive to create.

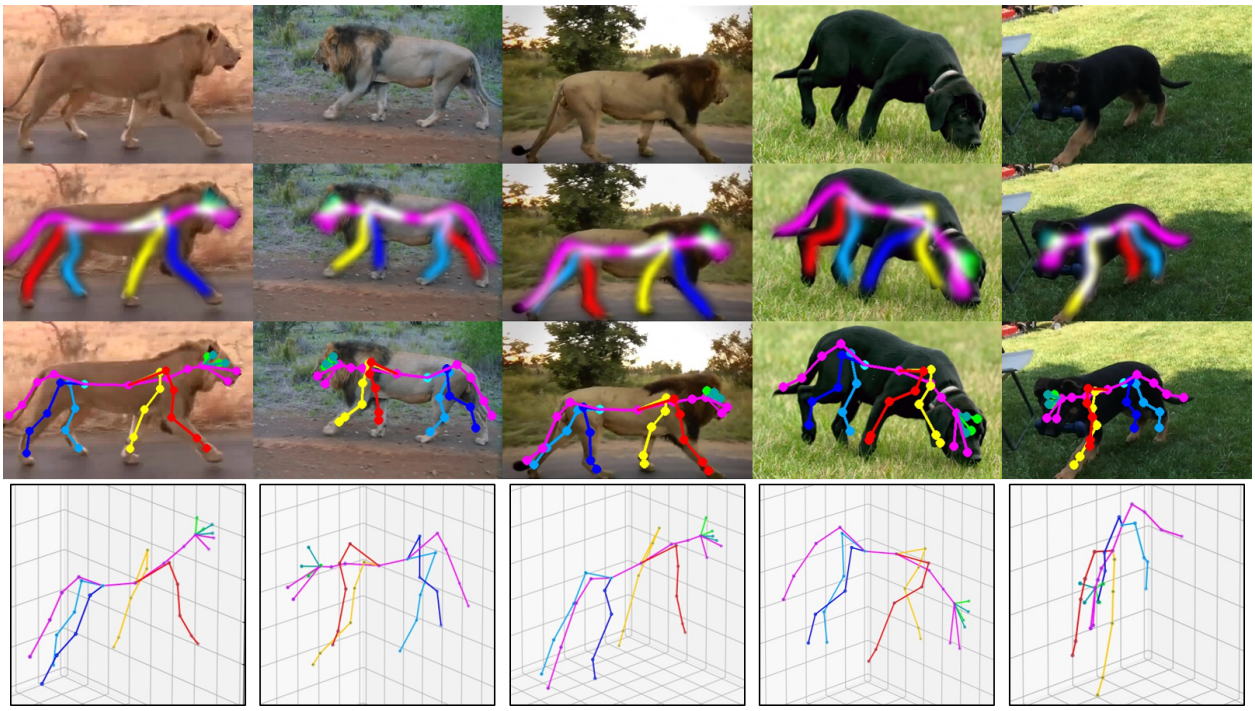

Figure 9: Estimated animal poses. Since the framework is generic it can be applied to other domains. We can successfully estimate poses for lions and dogs, that qualitatively outperform previous approaches such as Borer et al. [2021a].

8 Conclusion

We have shown that by extending the skeleton image representation to multiple channels in conjunction with mixing synthetic and unlabelled real data, we can reduce left and right ambiguity issues, avoid bad local minima and address the reality gap, without having to manually label data. Furthermore, we extended the analysis-by-synthesis pose estimation framework to predict 3D poses that overlap well with the input image thanks to the end-to-end training. And lastly, we presented an instance-specific refinement that allows us to considerably increase the performance on a completely new, unseen target video, without requiring any additional annotations for the target subject or environment.

References

Tomas Jakab, Ankush Gupta, Hakan Bilen, and Andrea Vedaldi. Self-supervised learning of interpretable keypoints from unlabelled videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

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, jul 2014.

Dominik Borer, Nihat Isik, Jakob Buhmann, and Martin Guay. Augmenting cats and dogs: Procedural texturing for generalized pet tracking. In *Proceedings of the 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - GRAPP*, 2021a. doi:10.5220/0010333701220132.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *ECCV*, 2014.

CMU. Cmu motion capture database. http://mocap.cs.cmu.edu/, 2001. Accessed: 2021-09-08.

Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3686–3693, 2014. doi:10.1109/CVPR.2014.471.

Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016.

Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *CoRR*, abs/1812.08008, 2018. URL `http://arxiv.org/abs/1812.08008`.

Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1653–1660, 2014. doi:10.1109/CVPR.2014.214.

Julieta Martinez, Rayat Hossain, Javier Romero, and James J. Little. A simple yet effective baseline for 3d human pose estimation. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2659–2668, 2017.

Gul Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017. doi:10.1109/cvpr.2017.492. URL `http://dx.doi.org/10.1109/CVPR.2017.492`.

Silvia Zuffi, Angjoo Kanazawa, David W. Jacobs, and Michael J. Black. 3d menagerie: Modeling the 3d shape and pose of animals. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5524–5532, 2017. doi:10.1109/CVPR.2017.586.

Silvia Zuffi, Angjoo Kanazawa, Tanya Berger-Wolf, and Michael J. Black. Three-d safari: Learning to estimate zebra pose, shape, and texture from images "in the wild". In *International Conference on Computer Vision*, October 2019.

Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Computer Vision and Pattern Regognition (CVPR)*, 2018.

Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3d human pose and shape from a single color image. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 459–468, 2018. doi:10.1109/CVPR.2018.00055.

Marc Habermann, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. Deepcap: Monocular human performance capture using weak supervision. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2020.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, October 2015.

Wei Yang, Wanli Ouyang, Xiaolong Wang, Jimmy Ren, Hongsheng Li, and Xiaogang Wang. 3d human pose estimation in the wild by adversarial learning. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5255–5264, 2018. doi:10.1109/CVPR.2018.00551.

Angjoo Kanazawa, David W. Jacobs, and Manmohan Chandraker. Warpnet: Weakly supervised matching for single-view reconstruction. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3253–3261, 2016. doi:10.1109/CVPR.2016.354.

James Thewlis, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object landmarks by factorized spatial embeddings. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

Tomas Jakab, Ankush Gupta, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object landmarks through conditional image generation. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL `https://proceedings.neurips.cc/paper/2018/file/1f36c15d6a3d18d52e8d493bc8187cb9-Paper.pdf`.

Yuting Zhang, Yijie Guo, Yixin Jin, Yijun Luo, Zhiyuan He, and Honglak Lee. Unsupervised discovery of object landmarks as structural representations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. URL `http://www.ytzhang.net/files/publications/2018-cvpr-lmdis-rep.pdf`.

Dominik Lorenz, Leonard Bereska, Timo Milbich, and Björn Ommer. Unsupervised part-based disentangling of object shape and appearance. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10947–10956, 2019. doi:10.1109/CVPR.2019.01121.

Jogendra Nath Kundu, Siddharth Seth, Varun Jampani, Mugalodi Rakesh, R. Venkatesh Babu, and Anirban Chakraborty. Self-supervised 3d human pose estimation via part guided novel image synthesis. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6151–6161, 2020. doi:10.1109/CVPR42600.2020.00619.

Luca Schmidtke, Athanasios Vlontzos, Simon Ellershaw, Anna Lukens, Tomoki Arichi, and Bernhard Kainz. Unsupervised human pose estimation through transforming shape templates. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2484–2494, June 2021.

Helge Rhodin, Mathieu Salzmann, and Pascal Fua. Unsupervised geometry-aware representation for 3d human pose estimation. In *Computer Vision – ECCV 2018*, pages 765–782, Cham, 2018. Springer International Publishing.

Ching-Hang Chen, Ambrish Tyagi, Amit Agrawal, Dylan Drover, Rohith MV, Stefan Stojanov, and James M. Rehg. Unsupervised 3d pose estimation with geometric self-supervision. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5707–5717, 2019. doi:10.1109/CVPR.2019.00586.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, 2014.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5967–5976, 2017. doi:10.1109/CVPR.2017.632.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.

Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8798–8807, 2018a. doi:10.1109/CVPR.2018.00917.

Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019a.

Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei Efros. Everybody dance now. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5932–5941, 2019. doi:10.1109/ICCV.2019.00603.

K. Aberman, M. Shi, J. Liao, D. Lischinski, B. Chen, and D. Cohen-Or. Deep video-based performance cloning. *Computer Graphics Forum*, 38(2):219–233, 2019. doi:https://doi.org/10.1111/cgf.13632. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.13632`.

Kripasindhu Sarkar, Lingjie Liu, Vladislav Golyanik, and Christian Theobalt. Humangan: A generative model of humans images. *CoRR*, abs/2103.06902, 2021. URL `https://arxiv.org/abs/2103.06902`.

Dominik Borer, Lu Yuhang, Laura Wülfroth, Jakob Buhmann, and Martin Guay. Rig-space neural rendering: Compressing the rendering of characters for previs, real-time animation and high-quality asset re-use. In *Proceedings of the 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - GRAPP*, 2021b. doi:10.5220/0010334503000307.

Joshua Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and P. Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 23–30, 2017.

Alexey Dosovitskiy and Thomas Brox. Generating images with perceptual similarity metrics based on deep networks. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL `https://proceedings.neurips.cc/paper/2016/file/371bce7dc83817b7893bcdeed13799b5-Paper.pdf`.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8798–8807, 2018b. doi:10.1109/CVPR.2018.00917.

Xudong Mao, Qing Li, Haoran Xie, Raymond Y.K. Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019b.

Cristian Sminchisescu Catalin Ionescu, Fuxin Li. Latent structured models for human pose estimation. In *International Conference on Computer Vision*, 2011.

Ruben Villegas, Jimei Yang, Yuliang Zou, Sungryull Sohn, Xunyu Lin, and Honglak Lee. Learning to generate long-term future via hierarchical prediction. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3560–3569. JMLR.org, 2017.

Praneet C. Bala, Benjamin R. Eisenreich, Seng Bum Michael Yoo, Benjamin Y. Hayden, Hyun Soo Park, and Jan Zimmermann. Automated markerless pose estimation in freely moving macaques with openmonkeystudio. *Nature Communications*, 11(1), December 2020. ISSN 2041-1723. doi:10.1038/s41467-020-18441-5.