Synthesizing Training Images for Boosting Human 3D Pose Estimation

Wenzheng Chen[1] Huan Wang[1] Yangyan Li[2] Hao Su[3] Zhenhua Wang[4]
Changhe Tu[1] Dani Lischinski[4] Daniel Cohen-Or[2] Baoquan Chen[1]

[1] Shandong University [2] Tel Aviv University [3] Stanford University [4] Hebrew University

1. Human3D+ Dataset for Evaluating 3D Pose Estimation

We show a sample of our Human3D+[1] images for 3D pose estimation in Figure 1[2]. In Figure 2, we show a visual comparison between the human skeletons generated by fine-tuning AlexNet on our synthetic images and the-state-of-the art method proposed in [5, 6] from images in Human3D+. It is clear that the model trained with our synthetic images performs much better qualitatively.

2. Synthetic Images for Boosting 2D Pose Estimation

2.1. Domain adaptation on 2D Pose Estimation

We have demonstrated that with domain adaptation, our synthetic data can benefit 3D pose estimation. In this section, we show that our synthetic images are also helpful in 2D pose estimation.

Similar to 3D pose estimation, in 2D pose estimation, we can make best use of synthetic images by training domain adaptation network. We show that our domain adaptation network trained with synthetic images and unannotated real images performs significant better than AlexNet trained with synthetic images. (see Figure 3 left). Meanwhile, we also show that our method performs better that the domain adaptation method proposed in [2] as we improve the training strategy to better adapt the features extracted from synthetic and real images.

As our domain adaptation network actually performs unsupervised domain adaptation, to better understand its effectiveness when supervised information is available, we evaluate its performance under different amount of annotated real images. Specifically, we train it with 100,000 synthetic images, 50,000 unannotated real images as unsupervised domain guidance, and 10, 100, 1,000, 10,000, or 100,000 annotated real images as supervised domain guidance. As a comparison, we also train AlexNet with the same amount

Figure 1. A sample of our Human3D+ images for 3D pose estimation. This dataset is captured in both indoor and outdoor scenes, containing images with various appearances and backgrounds.

[1]The new dataset, code and mdoel can be found at http://irc.cs.sdu.edu.cn/Deep3DPose/

[2]The sensors mounting strips are artificial, but necessary for accurate capturing. However, since such strips do not appear in Human3.6M or in our synthetic images, it is not harmful for the comparison fairness.

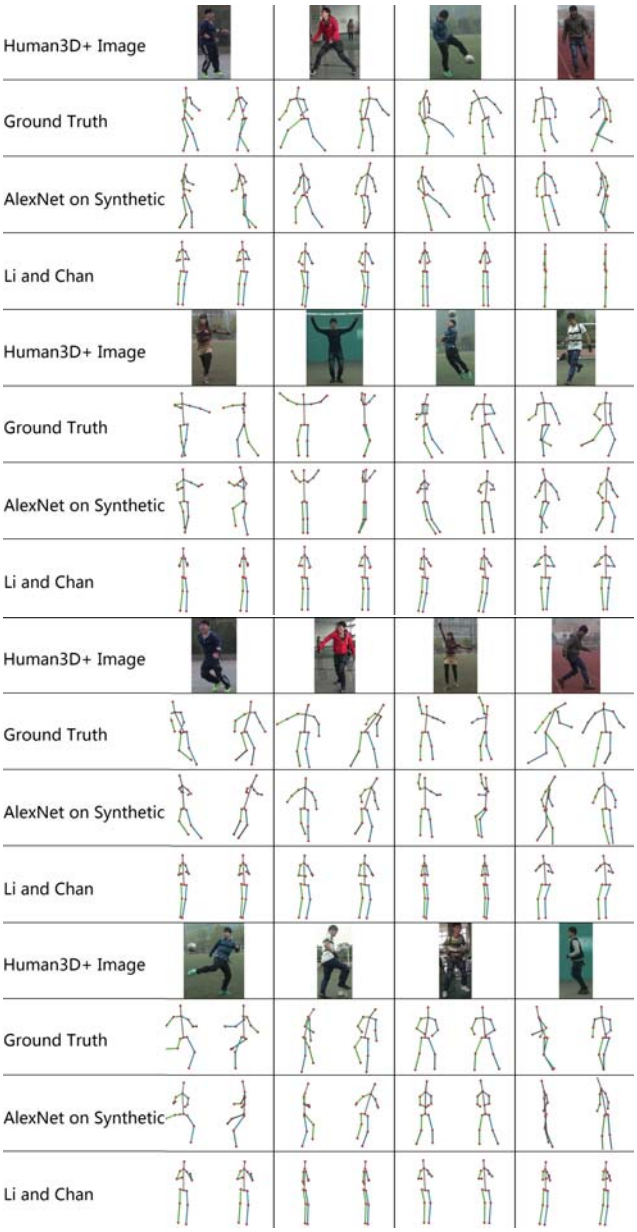

Figure 2. A visual comparison between the human skeletons generated by fine-tuning AlexNet on our synthetic images and the-state-of-the-art method proposed in [5, 6] from images in Human3D+.

of synthetic images and annotated real images. The result is shown in Figure 3 (right). It is obvious that our domain adaptation network works very effectively even if the supervised information is available. Unsurprisingly, the benefit brought by it is prominent especially when minimal amount of supervision is available.

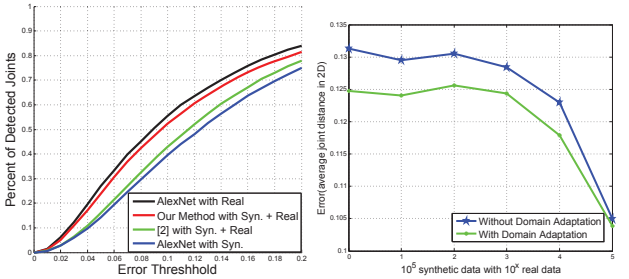

Figure 3. Left is the detection rate of joints with respect to different error thresholds. We can see that our domain adaptation network trained with synthetic images and unannotated real images performs significant better than AlexNet trained with synthetic images. Right shows the error rates when training with different amount of annotated realistic images.

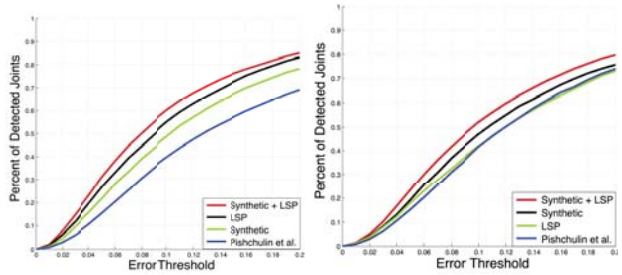

Figure 4. Performance of human 2D pose estimation CNNs trained on various datasets (LSP, Pishchulin et al., our synthetic and synthetic + LSP) evaluated on LSP (left) and MPI (right). We show that the performance of human 2D pose estimation CNNs can be consistenly improved by adding our synthetic images to real training images, which is probably due to that synthetic images can better cover human pose space.

2.2. Evaluation on 2D Pose Estimation

We have shown that our synthetic images serve better than state-of-the-art datasets for human 3D pose estimation in our main paper. We show that our synthetic images also boost human 2D pose estimation in this section.

Unlike 3D poses, 2D poses can be annotated by crowd sourcing, and there are several human images datasets with ground truth 2D pose annotations, e.g., LSP [3], MPI [1] and FLIC [8]. Since the images from these datasets are captured in less contrived settings, they can serve well for training, as well as evaluating, human 2D pose estimation models.

We show that the performance of human 2D pose estimation CNNs can be consistently improved by adding our synthetic images to real training images. More specifically, following deeppose [9], we fine-tune AlexNet [4] for human 2D pose estimation by changing the last fully connected layers from 1,000 to 30 (2D locations of 15), and transforming it to solve a regression task of the 2D joints with an Euclidean loss. We fine tune such a network on three datasets: 80,000 images augmented by applying mirror and rotation from 1,000 LSP training images (LSP), 250,000 images synthesized with our approach (Synthetic), and a mixture of the 1,000 LSP images with our synthetic images (Synthetic

+ LSP). We report their performance on LSP (1,000 testing images) and MPI (2,000 bounding box of full bodies) in Figure 4[3]. We compare the PDJ (Percentage of Detected Joints) error of the three models, where the horizontal axis corresponds to a threshold on the Euclidean distance between estimated poses and grountruth, and the vertical axis indicates the percentage of images that pass this threshold. We think the performance gain introduced by adding synthetic images into real training images is probably due to that synthetic images can better cover human pose space, which is a complementary information to real images.

We also compare our method to another human pose training image synthesis method proposed by Pishchulin et al [7]. To synthesize images, they first fit a 3D model to an image, which is not an automatic process, then texture the 3D model by the image. Finally, they deform the model and render the 3D model from the same view to generate new images. The manual 3D model fitting step and limitation on rendering from the same view prevent their method from scaling up well. Only 100 models were fit and used for synthesizing 3,000 images. We augmented their 3,000 images by mirroring and rotation to 120,000 images for training the aforementioned human 2D pose estimation CNN, and report its performance in Figure 4. We found their synthetic images are of higher realism, but limited in appearance, which is critical for training CNN models, thus do not perform as well as our synthetic images.

References

[1] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR*, June 2014. 4322

[2] Y. Ganin and V. Lempitsky. Unsupervised domain adaptation by backpropagation. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 1180–1189, 2015. 4321

[3] S. Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *BMVC*, 2010. doi:10.5244/C.24.12. 4322

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In P. Bartlett, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, editors, *NIPS*, pages 1106–1114. 2012. 4322

[5] S. Li and A. B. Chan. 3d human pose estimation from monocular images with deep convolutional neural network. In *ACCV*, pages 332–347. Springer, 2014. 4321, 4322

[6] S. Li, W. Zhang, and A. B. Chan. Maximum-margin structured learning with deep networks for 3d human pose estimation. In *ICCV*, pages 2848–2856, 2015. 4321, 4322

[7] L. Pishchulin, A. Jain, M. Andriluka, T. Thormählen, and B. Schiele. Articulated people detection and pose estimation: Reshaping the future. In *CVPR*, pages 3178–3185. IEEE, 2012. 4323

[8] J. Tompson, A. Jain, Y. Lecun, and C. Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. *NIPS*, 2014. 4322

[9] A. Toshev and C. Szegedy. DeepPose: Human pose estimation via deep neural networks. In *CVPR*, pages 1653–1660, 2014. 4322

[3] Evaluation on FLIC dataset is not performed, as it involves large portion of partial bodies, which is not the current focus of our synthesis approach.