[Supplementary Material] Learning Causal Representation for Training Cross-Domain Pose Estimator via Generative Interventions

Xiheng Zhang[1], Yongkang Wong[2], Xiaofei Wu[3], Juwei Lu[3], Mohan Kankanhalli[2],
Xiangdong Li[1*], Weidong Geng[1*]

[1]State Key Laboratory of CAD&CG, College of Computer Science and Technology, Zhejiang University
[2]School of Computing, National University of Singapore [3]Huawei Noah's Ark Laboratory

1. Model Architecture

The training procedure of the proposed method consists of several steps. Here we show the detail model architecture of each component in Table 1.

Encoder We construct 6 convolutional layers in the encoder E with 3×3 kernels, and the stride is fixed to be 2 to achieve spatial down-sampling instead of using deterministic spatial functions such as maxpooling. Each convolutional layer is followed by a batch normalization layer and a LeakyReLU activation layer. Then two fully-connected output layers (for mean and variance) are added to encoder, and will be used to compute the KL divergence loss [5] and sample latent variable (see [4] for details).

Decoder For decoder M, we use 6 convolutional layers with 3×3 kernels and set stride to be 1, and replace standard zero-padding with replication padding, i.e., feature map of an input is padded with the replication of the input boundary. For upsampling we use nearest neighbor method by a scale of 2 instead of fractional-strided convolutions. We also use batch normalization to help stabilize training and use LeakyReLU as the activation function.

Generator The feature generator g is built by the repetition of two blocks, each defined by a fully connected layer, a batch normalization layer, and a dropout layer, followed by a fully-connected layer with tanh activation functions.

Discriminator The discriminator $\mathcal{D}$ is a fully connected layer, with a sigmoid activate function as output layer. The extractor f is ResNet [2] or HRNet [13] backbone which takes in images patches cropped around the human or hand. And the predictor h takes the resulting feature map and upsamples it using three consecutive deconvolutional layers with batch normalization and ReLU. A 1-by-1 convolution is applied to the upsampled feature map to produce the volumetric heatmaps for each joint location.

Table 1: The architecture of variational autoencoder, feature generator and discriminator. We denote the 2D convolution as Conv, fully-connected layer as FC, LeakyReLU as LReLU and BN as batch normalization. S means the stride size of convolution and US represents upsampling. The output dimension of FC layer is denoted by d.

Module	Name	Layer	Output size
Encoder	Input		256×256×3
	Conv_0	3×3×8, S 2, BN, LReLU	128×128×8
	Conv_1	3×3×16, S 2, BN, LReLU	64×64×16
	Conv_2	3×3×32, S 2, BN, LReLU	32×32×32
	Conv_3	3×3×64, S 2, BN, LReLU	16×16×64
	Conv_4	3×3×128, S 2, BN, LReLU	8×8×128
	Conv_5	3×3×256, S 2, BN, LReLU	4×4×256
	FC_1&2	256 d	256, 256
Decoder	Input		256
	FC	4096 d	4×4×256
	Conv_0	US, 3×3×128, BN, LReLU	8×8×128
	Conv_1	US, 3×3×64, BN, LReLU	16×16×64
	Conv_2	US, 3×3×32, BN, LReLU	32×32×32
	Conv_3	US, 3×3×16, BN, LReLU	64×64×16
	Conv_4	US, 3×3×8, BN, LReLU	128×128×8
	Conv_5	US, 3×3×3	256×256×3
Generator	FC_0	1024 d, BN, ReLU, Dropout 0.5	1024
	FC_1	1024 d, BN, ReLU, Dropout 0.5	1024
	FC_2	256 d, BN, Tanh	256
Discriminator	FC_0	1024 d, ReLU	1024

2. Human Pose Datasets Summary

Human3.6M [3] dataset contains 3.6 million images featuring 11 actors performing 15 daily activities from 4 camera views. We follow the standard protocol and use subjects 1, 5, 6, 7 & 8 for training and subject 9 & 11 for evaluation. The evaluation is performed on every 64th frame of test set.

3DPW [16] dataset contains 60 clips of outdoor videos captured from a moving mobile phone and 17 IMUs attached to the subjects. We evaluate on the test set, which comprising 24 videos, using the 14 key joints that are common across both MS-COCO [6] and SMPL [8] skeletons.

MPI-INF-3DHP (3DHP) [9] dataset consists of both con-

Table 2: Human pose estimation results. The experiment is conducted on various `source→target` settings.

Learning Category	Methods	SURREAL→ 3DPW		SURREAL→ 3DHP		SURREAL→ H3.6M		SURREAL→ HumanEva	
		MPJPE↓	PAMPJPE↓	MPJPE↓	PAMPJPE↓	MPJPE↓	PAMPJPE↓	MPJPE↓	PAMPJPE↓
Conventional Learning	Source only	124.2	74.3	130.3	96.9	117.1	81.6	116.9	96.3
Domian Adaptation	DDC [14]	114.3	64.7	120.6	86.4	107.5	72.9	106.5	86.0
	DAN [7]	112.7	62.5	118.5	84.2	105.4	70.2	104.1	84.8
	DANN [1]	110.9	60.8	116.9	82.3	103.8	68.7	102.0	82.4
	Our method (SD + TD)	**103.2**	**54.4**	**107.1**	**76.3**	**94.9**	**59.5**	**94.7**	**77.8**
Domain Generalization	Wang *et al*. [17]	112.2	63.8	118.2	85.9	103.3	67.5	-	-
	Our method (SD + UD)	107.3	59.5	111.7	81.6	98.4	65.8	99.9	82.1
	Our method (SD + Multi-UDs)	**104.7**	**56.1**	**108.9**	**78.4**	**95.3**	**62.8**	**96.2**	**79.5**

strained indoor and complex outdoor scenes. It records 8 actors performing 8 activities from 14 camera views. On a 14-joint skeleton, we consider all the 8 actors and select sequences from 8 camera views as the training set. Evaluation is performed on the independent MPI-INF-3DHP test set.

SURREAL [15] is a synthetic person dataset generated with SMPL model [8]. It has 68036 videos of SMPL rendered humans moving on top of random backgrounds. The training, validation and test set consists of 55001, 507 and 12528 videos, respectively.

HumanEva [11] contains 7 calibrated video sequences that are synchronized with 3D body poses obtained from a motion capture system. The database contains 4 subjects performing a 6 common actions (e.g. walking, jogging, gesturing, etc.). The dataset contains training, validation and testing (with withheld ground truth) sets.

3. Hand Pose Datasets Summary

STB [12] dataset contains videos of a single person's left hand in 6 real world indoor environments. It has both 2D and 3D annotations of 21 joints for 18000 stereo pairs. We follow [12] and divide the dataset into a training set of 15000 images and an evaluation set of 3000 images.

RHD [19] is a synthetic dataset built upon 20 different characters performing 39 actions. It provides 41,258 images for training and 2,728 images for evaluation.

FreiHAND [20] is a 3D hand pose dataset which records different hand actions performed by 32 people. It contains 32,560 training samples and 3960 samples for evaluation.

Panoptic (PAN) [19] dataset was recorded using a multi-view capture setup with 10 RGB-D sensors, 480 VGA and 31 HD cameras. We select 171204_pose3 as evaluation set and use the remaining 11 sequences for training.

GANerated (GAN) [10] dataset has 330K synthetic images of hands, sometimes holding an object, in a random background. The images were made more realistic by extending CycleGAN [18].

4. Evaluation on Human Pose Estimation Task

In this section, we provide additional evaluation results on human pose estimation task. We select SURREAL as the source dataset where 3DPW, 3DHP, Human3.6M and HumanEva are used in turn as target dataset. The naïve baseline model is trained on source dataset only and directly tested on the target dataset without any adaptation. Table 2 shows the results of several baselines and our proposed method. For the domain adaptation setting, our proposed approach outperforms DDC [14], DAN [7], DANN [1] with a significant margin on both MPJPE and PAMPJPE.

We also evaluate on the domain generalization setting where there is no access to the target domain data. When only using one unconstrained dataset, our method (SD + UD) reduces MPJPE by an average of 5.43 mm on three target datasets when compared with Wang *et al.* [17]. In addition, when using multiple unconstrained datasets, our method (SD + Multi-UDs) can even reach a competitive performance against the domain adaptation model, *i.e.* our method (SD + TD).

5. Ablation Study

Dimension of Generated Features Here, we examine how the dimensionality of generated features influence the performance of our method. Table 3 show the experimental results of different feature dimension used in training. It can be seen that choosing a dimension of 256 is better than the others. When we increase the dimension from 256 to 1024, the performance of our proposed model declines slowly. As the dimension of features increases, the parameters of the model become more complex. Hence the bigger the dimension of features, the more the chances of overfitting. Choosing a relatively small dimension could help the model avoid overfitting to the training data.

Visualization of the Generated Features Here, we visualize the generated features from 6 randomly selected GT poses with t-SNE algorithm. In Figure 1, dots with the same color are features generated from the same GT pose but with different random noise. The features generated from

Table 3: Human pose estimation performance of different feature dimension on 3DPW dataset.

Feature Dimension	MPJPE↓	PAMPJPE↓
128	95.8	65.7
256	**94.7**	**63.9**
512	96.3	64.4
1024	97.2	67.5

the same pose clump together as a cluster, while those from different poses are far away. This empirically implies the pose information is properly embedded. Furthermore, for similar pose (e.g., red & blue and orange & green), the generator can still generate distinguishable features that are not fully overlapped.

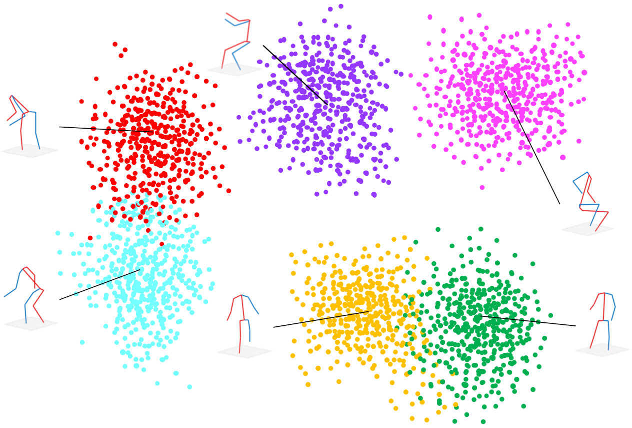

Figure 1: Visualization of the distribution of generated features from 6 random GT poses with t-SNE algorithm.

References

[1] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, pages 1180–1189, 2015.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[3] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *TPAMI*, 36(7):1325–1339, 2014.

[4] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014.

[5] Solomon Kullback. Letter to the editor: The kullback-leibler distance. *The American Statistician*, 1987.

[6] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755, 2014.

[7] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *ICML*, pages 97–105, 2015.

[8] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *TOG*, 34(6):1–16, 2015.

[9] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3D human pose estimation in the wild using improved cnn supervision. In *3DV*, pages 506–516, 2017.

[10] Franziska Mueller, Florian Bernard, Oleksandr Sotnychenko, Dushyant Mehta, Srinath Sridhar, Dan Casas, and Christian Theobalt. Ganerated hands for real-time 3D hand tracking from monocular rgb. In *CVPR*, pages 49–59, 2018.

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

[12] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *CVPR*, pages 1145–1153, 2017.

[13] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, pages 5693–5703, 2019.

[14] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014.

[15] Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *CVPR*, pages 109–117, 2017.

[16] Timo von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3D human pose in the wild using IMUs and a moving camera. In *ECCV*, volume 11214 of *Lecture Notes in Computer Science*, pages 614–631, 2018.

[17] Zhe Wang, Daeyun Shin, and Charless C Fowlkes. Predicting camera viewpoint improves cross-dataset generalization for 3D human pose estimation. In *ECCV Workshops*, volume 12536 of *Lecture Notes in Computer Science*, pages 523–540, 2020.

[18] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, pages 2223–2232, 2017.

[19] Christian Zimmermann and Thomas Brox. Learning to estimate 3D hand pose from single RGB images. In *ICCV*, pages 4903–4911, 2017.

[20] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. FreiHAND: A dataset for markerless capture of hand pose and shape from single RGB images. In *ICCV*, pages 813–822, 2019.