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

Abstract

3D pose estimation has attracted increasing attention with the availability of high-quality benchmark datasets. However, prior works show that deep learning models tend to learn spurious correlations, which fail to generalize beyond the specific dataset they are trained on. In this work, we take a step towards training robust models for cross-domain pose estimation task, which brings together ideas from causal representation learning and generative adversarial networks. Specifically, this paper introduces a novel framework for causal representation learning which explicitly exploits the causal structure of the task. We consider changing domain as interventions on images under the data-generation process and steer the generative model to produce counterfactual features. This help the model learn transferable and causal relations across different domains. Our framework is able to learn with various types of unlabeled datasets. We demonstrate the efficacy of our proposed method on both human and hand pose estimation task. The experiment results show the proposed approach achieves state-of-the-art performance on most datasets for both domain adaptation and domain generalization settings.

1. Introduction

3D pose estimation has been attracting increasing attention due to its numerous applications in human-computer interaction, action recognition and privacy preservation application [9, 59, 68]. In recent years, the deep learning models have achieved tremendous improvement with advance in model architecture [7, 17, 57], novel loss functions [24, 35], and availability of quality datasets [6, 36, 21]. Despite its success, existing methods still struggle to generalize beyond the domain of training data, where a well-trained model is unable to detect precise joints locations in unfamiliar subjects or unseen views (*i.e.* cross-domain pose estimation).

*Corresponding authors

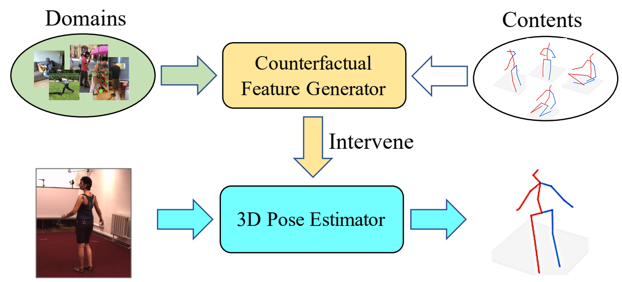

Figure 1: Overview of the training process of robust pose estimator with generative interventions. Given a set of domains and content, we train a generator that produce counterfactual features to intervene the training of an estimator.

The deficiency on cross-domain pose estimation can be attributed to dataset biases [61] or shortcut learning [11], which means that deep learning models are prone to learn dataset-dependent spurious correlations based on statistical associations [1, 2, 4, 20, 48]. This characteristic becomes problematic when the correlations are not consistent across domains. For 3D pose estimation task, an example of spurious correlation could be the connection between the appearance of clothes/skin and joints. Generally, this is not a problem during the inference stage as long as the data follows the same distribution. However, the test samples could comprise individuals with skin color or clothes that are different from the training dataset. Hence, the trained model's performance might not be as good as expected.

Prior works show that generalizing beyond training domain requires a model to learn not only the statistical associations between variables, but also the underlying causal relations [52]. Causal relations reflect the fundamental data-generating mechanism, which tends to be universal and invariant across different domains [49], and provides the most transferable and confident information to unseen domains. For example, composing a shot on photography involves both content (*e.g.* person, object, *etc.*) and a specific domain (*e.g*. background, viewpoint, *etc.*). Even though the domain may differ, the photo's semantics would remain consistent as long as the content is unchanged. The goal of causal representation learning is to learn a representation exposing the causal relation which is invariant under different interventions. This allows a learning framework to train predictive models that are robust against the changes in domain that naturally occur in the real world.

In this paper, we propose a novel method for learning causal representations, which is subsequently used to train a robust model for cross-domain pose estimation task. The proposed method is based on the observation that the causal generative process of an image, which assumes the data is constructed from a content variable and a domain variable, is domain- or dataset-invariant. Building on prior work [12, 18], we consider changing the domain variable as an intervention on the images. We then do such interventions by steering the generative models to produce *counterfactual features* from a specified content and random noise. Finally, by enforcing similarity between the distribution of representations learned with different interventions, the model can learn transferable and causal relations across different domains. An overview of the pose estimator training with the counterfactual representation is shown in Figure 1.

The main contributions of our work are as follows:

- We propose a novel framework for causal representation learning to generate out-of-distribution features. We explicitly exploit the causal structure of the task and show how to learn causal representations by steering the generative model to produce counterfactual features, which simulates domain interventions on images.
- We demonstrate the effectiveness of the counterfactual feature generator by utilizing the generated features to train models for pose estimation task. Not only can our method enhance the cross-domain pose estimation performance (*i.e*. train with both source domain data and unlabeled target domain data), but also generalize well to domain generalization setting (*i.e*. train with both source domain data and unlabeled unconstrained dataset).
- We conduct experiments on both human pose and hand pose estimation task. The ablation studies examine various components of the proposed framework, as well as the impact of different mixture of training datasets. We also discuss why increasing source dataset and intervention can improve performance.

2. Related Work

3D Pose Estimation With the recent advances of deep learning, there are significant improvements in 3D human pose estimation [5, 34, 41, 50, 60] and 3D hand pose estimation [45, 46, 62, 76]. A number of works focused on the cross-domain scenario. Zhou *et al*. [75] proposed a weakly-supervised transfer learning method with 3D geometric constraint, which uses mixed 2D and 3D labels from indoor-datasets and in-the-wild datasets. Habibie *et al*. [15] proposed a new disentangled hidden space encoding of explicit 2D and 3D features for monocular 3D human pose estimation that shows high accuracy and generalizes well to in-the-wild scenes. Zhang *et al*. [73] proposed a domain adaptation framework with unsupervised knowledge transfer, which aims at leveraging the knowledge in multi-modality data of the easy-to-get synthetic depth datasets to better train a pose estimator on the real-world datasets. Zimmermann *et al*. [77] analyzed cross-dataset generalization when training on existing hand pose estimation datasets. They also introduced a large-scale multi-view hand dataset with both 3D hand pose and shape annotations. Wang *et al*. [67] carried out a systematic study of the diversity and biases present in specific datasets, and its effect on cross-dataset generalization across five human pose datasets. Zhao *et al*. [74] introduced an end-to-end scheme for cross-modal knowledge generalization to transfer cross-modal knowledge between source and target hand pose datasets where superior modalities are missing. Baek *et al*. [3] proposed an end-to-end trainable pipeline that adapts hand-object domain to single hand-only domain, where hand-object images are translated to segmented and de-occluded hand-only images.

Causal Representation Learning Traditional causal discovery and reasoning assume that the units are random variables connected by a causal graph. However, real world observations are usually unstructured, *e.g*. objects in a given image [38]. Hence, the emerging field of causal representation learning strives to learn these variables from data. Previous works have attempted to combine causal structural modeling and representation learning. Shalit *et al*. [54] gave a new theoretical analysis and family of algorithms for predicting individual treatment effect from observational data. The algorithms learn a balanced representation such that the induced treated and control distributions are similar. As described in the perfect match approach [53], this model can also be extended to any number of treatments by augmenting samples within a mini-batch with their propensity-matched nearest neighbours. Following this idea, Johansson *et al*. [23] brought together shift-invariant representation learning and re-weighting methods. Hassanpour and Greiner [16] presented a context-aware weighting scheme based on the importance sampling technique to alleviate the selection bias problem. Yao *et al*. [70, 71] proposed a local similarity preserved individual treatment effect estimation method to preserves local similarity and balances data distributions simultaneously.

Domain Adaptation/Generalization To mitigate dataset bias or domain shift in cross-domain scenario, domain adaptation/generalization has attracted a lot of attention. Domain adaptation methods aim to reduce domain shift by explicitly align the source and target distribution [10, 37, 63, 64]. Domain generalization relates to domain adaptation in that it aims to improve target domain's performance, rather than source domain. However, it considers the case where the model learns to generalize from a set of source domains without accessing target domain samples during training phrase [31, 32, 33, 44]. Recently, leveraging causality as a notion of invariant prediction has emerged as an important operational concept in causal inference. Mitrovic *et al*. [43] analyzed self-supervised representation learning with causal framework. They proposed a self-supervised objective that enforces invariant prediction of proxy targets across augmentations through an invariance regularizer which yields improved generalization guarantees. Mao *et al*. [39] learn discriminative visual models that are consistent with causal structures to enables robust generalization. By steering generative models to construct interventions, they randomize many features without being affected by confounding factors. Sauer and Geiger [51] proposed to decompose the image generation process into independent causal mechanisms, which disentangle object shape, object texture and background, for generating counterfactual images that improve out-of-distribution robustness.

Key Novelties Our work has three key differences from the above works. Firstly, unlike [43], we consider domains as interventions rather than simple image transformations. The variation of domain is more general than the image transformations and naturally occur in the real world. Secondly, unlike [39, 51], our method proposes to generate *counterfactual features* rather than counterfactual images. Most GAN-based image generation methods are unstable and usually suffer from ghosting effect. In pose estimation task, locating each joint is heavily reliant on the distinct details of human body. Low-quality images with uncertain ground-truth could degrade the model performance. Alternatively, we propose to train a feature generator which directly produces *counterfactual features* from a specified content and random noise. Lastly and most importantly, our proposed framework additionally enforces similarity between the distribution of representations learned with different interventions, on the basis of maximizing the training set margin. The trade-off between different objectives helps our model to learn transferable and causal relations across domains.

3. Preliminary

This section first overviews the structural causal model and causal inference problem. Then, it shows a causal view of data generation process in CV tasks and formulates the problem of causal inference for domain adaptation.

3.1. Structural Causal Models & Causal Inference

The Structural Causal Models (SCMs) [47] consider a set of variables $X_1, ..., X_n$ associated with the vertices of a directed acyclic graph. We assume that each variable is the result of an assignment using a deterministic function f_i depending on X_i's parents in the graph (denoted by $\mathbf{PA}_i$) and an unexplained random variable U_i, *i.e*.

$$X_i := f_i(\mathbf{PA}_i, U_i), (i = 1, ..., n), \tag{1}$$

Directed edges in the graph represent direct causation. In SCM, *intervention* is formalized as operations that modify a subset of assignment in Eq. 1, *e.g*. changing U_i, setting f_i to a constant, or changing the functional form of f_i (and thus changing the dependency of X_i on its parents) [47].

The problem of causal inference is to estimate the outcome changes if a different interventions had been applied [69]. For example, suppose two treatments, *i.e*. medicine A and B, are available to patients. Given that the recovery rate for patients who took medicine A is 70%. Would they have a higher recovery rate had they received another medicine? Such questions are termed counterfactual questions [30]. Formally, let $\mathcal{T}$ be the set of potential interventions, $\mathcal{X}$ the set of units, and $\mathcal{Y}$ the set of potential outcomes. In the case of binary action set $\mathcal{T} = \{0, 1\}$, the observed samples consist of set $\hat{P}^{\mathrm{F}} = (x_i, t_i)_{i=1}^n$ and the counterfactual samples consist of set $\hat{P}^{\mathrm{CF}} = (x_i, 1 - t_i)_{i=1}^n$. Here, set $\hat{P}^{\mathrm{F}} \sim P^{\mathrm{F}}$ is the empirical observed distribution and set $\hat{P}^{\mathrm{CF}} \sim P^{\mathrm{CF}}$ is the empirical counterfactual distribution. As only one potential outcome could be observed, we define the observed outcomes as $y^{\mathrm{F}}(x)$ and the unobserved outcomes as counterfactual outcomes $y^{\mathrm{CF}}(x)$.

3.2. A Causal View of Data-Generation Process

Consider a generic computer vision task where a model is trained with curated image data, a basic assumption in causal inference is that the test data may be sampled from a different distribution but comprises the same causal mechanisms as in the training dataset [49]. For example, composing a shot on photography involves both content (*e.g*. person, object, *etc*.) and a specific domain (*e.g*. background, viewpoint and camera setting). Even though the domain may differ, the photo's semantics would remain consistent as long as the content is unchanged.

Figure 2 shows a causal graph that describes a data-generation process. The images are caused by both the content variable C and domain variable D, as shown by the

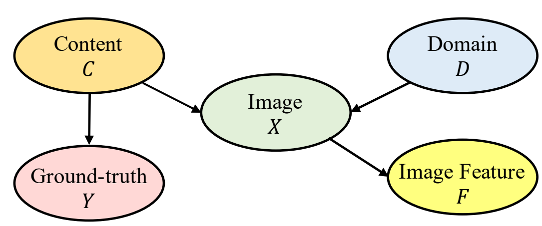

Figure 2: Structural causal graph for computer vision task.

two incoming arrows to X. The arrow from C to Y indicates the ground-truth Y is conditioned on content variable C. In addition, we introduce a node of image features F extracted by an encoder. If we consider the images X as the set of unit $\mathcal{X}$, changing the domain D can be seen as an intervention on a image x_i: for each observed sample $\{x_i, d_i\}$, there is a set of (unobserved) counterfactual samples $\{x_i, d_j\}$ where $d_i \neq d_j$. And the ground-truth consists of the set of potential outcomes Y. Let D_s and D_t represent source and target domain. Then the set of interventions is $\mathcal{T} = \{D_s, D_t\}$. Specifically, $P^{F}(X, D) = P(X) \cdot P(D_s|X)$ and $P^{CF}(X, D) = P(X) \cdot P(D_t|X)$. The difference between the observed and counterfactual samples lies precisely in the intervention assignment mechanism, $P(D|X)$ [22]. X and D are not independent according to the causal graph. As a result, P^{CF} will generally be different from P^{F}.

4. Method

The essence of the proposed method is to learn causal representations exposing the causal relation that is invariant under different interventions. Here, we will first delineate the proposed approach under the domain adaptation setting, and then extend it to the domain generalization scenario.

4.1. Learning with Causal Representation

Taking into account that a robust model must learn to generalize from source domain (observed) distribution to the target domain (counterfactual) distribution, we propose to learn causal representations which trade-off between three objectives: (1) Enabling low-error prediction over the observed representations; (2) Enabling low-error prediction over the counterfactual representations; (3) The distributions of different intervention populations are similar.

The proposed framework that simultaneously accomplishes these objectives is depicted in Figure 3. Specifically, it contains two branches: In the observed representation branch, there is a feature extractor f, which takes the images from source domain as input and yields representations over the observed distribution. The counterfactual representation branch consists of a feature generator g which produces counterfactual features from a ground-truth pose and random noise. After obtaining both observed and counterfactual representations, they are fed into a predictor h to obtain volumetric heatmaps which can be converted to 3D pose by applying soft-argmax activating function.

In practice, we have access to neither the counterfactual samples as opposed to the observed samples (changing domain would cause the variation of image) nor the potential outcomes of such counterfactual samples. In previous work, [39, 51] proposed to utilize generative models to produce counterfactual images. The GAN-based image generation methods are unstable and usually suffer from ghosting effect for unseen domains. In the pose estimation task, locat-

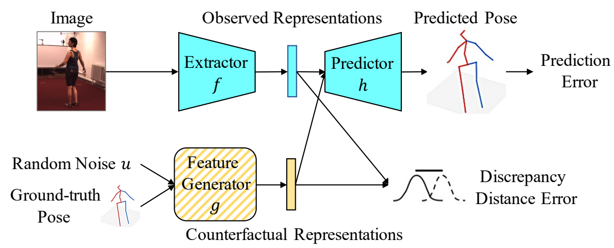

Figure 3: An overview of the proposed domain adaptation learning framework with novel counterfactual feature.

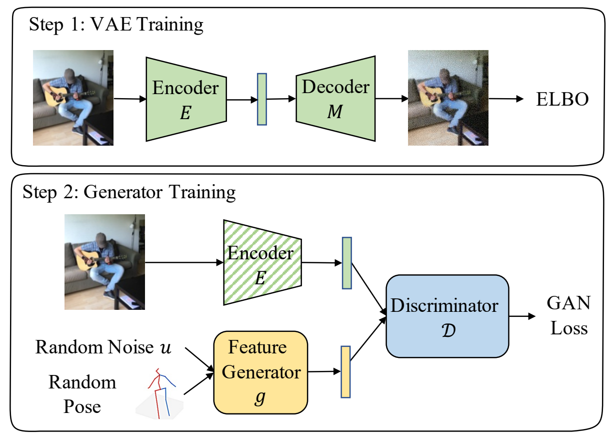

Figure 4: An overview of the two steps training process of the counterfactual feature generator.

ing each joint is heavily reliant on the distinct visual details of human body, and fuzzy images with uncertain ground-truth could worsen the model performance. As an alternative, we propose to train a feature generator g which directly produces counterfactual features instead of images.

Under the proposed framework, we could accomplish the first and the second objectives by empirical risk minimization over both the observed and counterfactual distributions. In addition, we manage the third objective by enforcing the similarity between the distribution of different intervention groups in the representation space. Specifically, we minimize the discrepancy distance between the observed and counterfactual representations, which encourages the model to learn underlying invariances which generalize from the observed distribution to counterfactual distributions.

4.2. Counterfactual Feature Generator

Figure 4 illustrates the two steps training procedure of the proposed generator. The first step is to train a Variational Autoencoder (VAE) $E \circ M$. The encoder E takes image x as input and encodes it into a latent embedding $z = E(x) \sim q(z|x)$, while the decoder M learns to reconstruct the image from the latent embedding, *i.e.* $\hat{x} = M(z) \sim p(x|z)$. The objective is defined as the minimization of the Evidence Lower Bound (ELBO) [27]:

$$\min_{\theta_E,\theta_M} \mathcal{L} = -\mathbb{E}_{q(z|x)}[\log p(x|z)] + \mathrm{KL}\big(q(z|x)\|p(z)\big) \quad (2)$$

where θ_E and θ_M are the parameters of E and M, respectively. KL() is the Kullback–Leibler divergence [29]. Once the VAE is appropriately trained, we can encode an image to the latent embedding space. The image's latent embedding should contain all the information required to construct it.

In the second step, we propose to learn a feature generator g with the help of the pretrained VAE in an adversarial manner. Specifically, the generator g takes a noise vector u sampled from a spherical Gaussian distribution $p(u)$ and a pose label y as input. Given the encoder E of VAE, we train the generator g to produce features, of which the distribution resembles that of the latent embedding from the encoder as much as possible, such that a discriminator $\mathcal{D}$ cannot reliably distinguish them. Based on the Least Squares GAN [40], the following min-max game is defined:

$$\min_{\theta_g}\max_{\theta_{\mathcal{D}}} \mathcal{L} = \mathbb{E}_{(u,y)\sim(p(u),p(y))}\|\mathcal{D}\big(g(u,y)\big)-1\|^2 + \mathbb{E}_{x\sim p(x)}\|\mathcal{D}\big(E(x)\big)\|^2 \quad (3)$$

where θ_g (θ_D) is the parameters of generator (discriminator). $p(x)$ and $p(y)$ represent the distribution of input image and input pose label, respectively. The parameters in E are kept frozen when training the generator g. Once the feature generator g is appropriately trained, it is used to obtain features from a random pose. The distribution of the features should be similar to the distribution of latent embedding.

4.3. The Overall Training Procedure

The overall training pipeline of our proposed framework can be described as follows: (1) Training a VAE $E \circ M$ over the images from target domain, (2) Training a counterfactual feature generator g with the help of the encoder E while the parameters of E are kept frozen. The encoder E takes images from the target domain as input, while the feature generator g takes random poses from source domain as input. (3) Training the feature extractor f and predictor h with the help of the feature generator g while the parameters of g are kept frozen (as shown in Figure 3). The feature extractor f and feature generator g respectively takes image and pose from source domain as input.

The overall objective in step (3) is defined as follow:

$$\min_{\theta_f,\theta_h} \mathcal{L} = \mathbb{E}_{(x,y,u)\sim(p(x),p(y),p(u))}\mathcal{L}_{\mathrm{F}}\big(h(f(x)),y\big) + \lambda_1\mathcal{L}_{\mathrm{CF}}\big(h(g(u,y)),y\big) + \lambda_2\mathcal{L}_{\mathrm{dist}}\big(f(x),g(u,y)\big) \quad (4)$$

where λ_1 and λ_2 control the strength of the imbalance penalties. The loss term $\mathcal{L}_{\mathrm{F}}$ and $\mathcal{L}_{\mathrm{CF}}$ stand for the prediction error over observed and counterfactual distributions, respectively. We use smooth-l_1 distance to compute the error between ground-truth pose and the predicted pose. For discrepancy distance loss $\mathcal{L}_{\mathrm{dist}}$, we select Maximum Mean Discrepancy [14], KL divergence [29] and naïve l_2 for experiments. The performance of each loss was reported in Section 5.4.

4.4. Extension to Domain Generalization

For the domain generalization setting, where the model learns to generalize from a set of domains (*i.e.* without target domain samples), the proposed framework can be extended from binary interventions to multiple interventions. Let the total number of interventions be K. Only one intervention can be provided to an individual i. Accordingly, the observed outcome of a unit x_i under the k-th intervention is given by y^{F}_{ik}. The counterfactuals are defined under the $K-1$ alternate interventions which are unobserved. Specifically, we select one dataset as source domain, and consider the others as interventions. Then we can train $K-1$ counterfactual feature generators for the source dataset. To train the model, the first and second objectives remain the same as binary interventions, while the third objective changes to a sum of pair-wise discrepancy distance error between the observed and each counterfactual representations.

5. Experiments

To validate the effectiveness of our proposed method, we conduct three kinds of experimental settings on both human and hand pose estimation task. An overview of the experiment settings is shown in Table 1. Firstly, the conventional learning trains model only on source domain (SD) data and test on the target domain (TD) data. Secondly, in domain adaptation, both SD and TD data are available for training, and TD is used for validation. The label of TD data is not available during training. Thus, it is considered as unsupervised domain adaptation. At last, different from domain adaptation, the TD data is inaccessible during the training for domain generalization scenario. Instead, we explore the rich data from source domain, as well as introduce unconstrained domain (UD) data as a supplement to SD data during training. The UD data include action recognition datasets [25, 28, 58] and general image datasets [6, 8, 36].

5.1. Datasets and Evaluation Metrics

Human Pose Estimation Task We evaluate on five human pose datasets, namely Human3.6M [21], 3DPW [66],

Table 1: Overview of the experiment settings. SD: Source domain; TD: Target domain; UD: Unconstrained domain.

Task	Training Set	Test Set
Conventional Learning	SD	TD
Domain Adaptation	SD + TD (w/o label)	TD
Domain Generalization	SD + UD	TD

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

Learning Category	Methods	H3.6M→ 3DPW		H3.6M→ 3DHP		H3.6M→ SURREAL		H3.6M→ HumanEva	
		MPJPE↓	PAMPJPE↓	MPJPE↓	PAMPJPE↓	MPJPE↓	PAMPJPE↓	MPJPE↓	PAMPJPE↓
Conventional Learning	Source only	118.7	78.0	121.8	98.5	128.6	86.5	91.6	77.2
Domian Adaptation	DDC [64]	110.4	75.3	115.6	91.5	117.5	80.1	83.8	64.9
	DAN [37]	107.5	73.2	109.5	89.2	114.2	78.4	78.5	62.7
	DANN [10]	106.3	71.1	107.9	88.0	113.6	77.2	76.3	60.8
	ISO [72]	-	70.8	-	**75.8**	-	-	-	-
	Our method (SD + TD)	**94.7**	**63.9**	**99.3**	81.5	**103.3**	**69.1**	**69.2**	**53.5**
Domain Generalization	Wang *et al.* [67]	109.5	68.3	111.9	89.0	114.0	75.9	-	-
	Our method (SD + UD)	97.5	66.4	102.6	86.4	108.8	74.3	75.6	58.1
	Our method (SD + Multi-UDs)	**94.9**	**64.2**	**101.6**	**83.7**	**105.6**	**72.5**	**70.7**	**54.4**

MPI-INF-3DHP (3DHP) [42], SURREAL [65] and HumanEva [55]. The details can be found in the supplementary material. We adopt two commonly used metrics, the Mean Per Joint Position Error (MPJPE) to compute the mean Euclidean distance between ground-truth and predicted pose, whereas the Procrustes Aligned Mean Per Joint Position Error (PAMPJPE) computes MPJPE based on the predicted pose aligned to ground-truth by the Procrustes method [13].

Hand Pose Estimation Task We evaluate on five hand pose dataset, namely STB [56], RHD [76], FreiHAND [77], Panoptic (PAN) [76] and GANerated (GAN) [45]. The details can be found in the supplementary material. We report results using two metrics. The mean end-point-error (EPE) is defined as the average Euclidean distance between predicted and ground-truth keypoints. The area under the curve (AUC) on the percentage of correct keypoints (PCK) score. PCK is the percentage of predicted joints that fall within the given threshold distance with respect to the ground-truth.

5.2. Implementation Details

As different human pose datasets have diverse joint configuration, we follow [67] to select a subset of 14 common joints to eliminate the bias introduced by a different number of joints during training. We normalize the z value from $(-z_{\text{max}}, +z_{\text{max}})$ to (0, 63) for integral regression. z_{max} is set to 2400 mm based on all datasets. Similarly, we follow [77] to select 20 common joints on hand pose datasets.

We use PyTorch to implement our network. ResNet and HRNet are initialized using the pretrained weights on ImageNet dataset [6]. We use Adam optimizer [26] with a mini-batch size of 128. The initial learning rate is set to 1×10^{-3} and reduced by a factor of 10 at the 170th epoch. We use 256×256 and 384×288 as the input size of ResNet and HRNet, respectively. The data augmentation scheme includes random rotation ($[-45°, 45°]$), random scale ($[0.65, 1.35]$), and flipping. The variational autoencoder is based on the structure in [19]. The detail model architecture of each component can be found in supplementary material.

5.3. Results on Human/Hand Pose Estimation

Human Pose Estimation In this section, we validate the efficacy of the proposed method on the human pose estimation task. In all experiments, we select Human3.6M as the source dataset where 3DPW, 3DHP, SURREAL 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 [64], DAN [37], DANN [10] with a significant improvement on both MPJPE and PAMPJPE. Specifically, our method improves the PAMPJPE metric by 6.9 mm on 3DPW, 8.1 mm on SURREAL and 7.3 mm on HumanEva.

We also evaluate for the domain generalization setting where there is no access to the target domain data. Here, we use common action recognition datasets as supplement training data, including UCF101 [58], HMDB [28] & Kinetics [25]. When only using one unconstrained dataset, *i.e.* Kinetics, our method (SD + UD) reduces MPJPE by an average of 8.83 mm on three target datasets when compared with Wang *et al.* [67]. 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).

Hand Pose Estimation This section discusses model performance on the hand pose estimation task. In all experiments, we select FreiHAND as the source dataset where STB, RHD, PAN and GAN are used in turn as target dataset. The results are shown in Table 3. Overall, the improvement trend is similar as that of the human pose estimation task. For the domain adaptation setting, our method (SD + TD) brings a significant improvement on both EPE and AUC over the state-of-the-art methods. For domain generalization setting, our method (SD + UD) also improves AUC by an average of 0.5275 when compared with [77]. When using multiple unconstrained datasets, our

Table 3: Hand pose estimation results. The experiment is conducted on various `source→target` settings.

Learning Category	Methods	FreiHAND→ STB		FreiHAND→ RHD		FreiHAND→ PAN		FreiHAND→ GAN	
		EPE ↓	AUC ↑	EPE ↓	AUC ↑	EPE ↓	AUC ↑	EPE ↓	AUC ↑
Conventional Learning	Source only	36.1	0.433	48.3	0.287	35.6	0.453	59.4	0.156
Domain Adaptation	DDC [64]	34.5	0.462	44.6	0.355	32.5	0.525	57.3	0.175
	DAN [37]	32.7	0.514	40.5	0.387	32.1	0.548	54.9	0.201
	DANN [10]	30.9	0.576	38.0	0.411	31.8	0.553	53.6	0.224
	Our method (SD + TD)	**22.4**	**0.619**	**35.4**	**0.458**	**22.9**	**0.613**	**49.5**	**0.278**
Domain Generalization	Zimmermann *et al*. [77]	-	0.52	-	0.399	-	0.562	-	0.217
	Our method (SD + UD)	29.3	0.584	37.6	0.423	31.3	0.572	52.7	0.235
	Our method (SD + Multi-UDs)	**24.2**	**0.603**	**35.7**	**0.444**	**28.6**	**0.596**	**50.6**	**0.266**

Table 4: Human pose estimation performance of various backbone architectures on 3DPW dataset.

Method	Backbone	MPJPE↓	PAMPJPE↓
Source only	ResNet-18	122.4	83.1
	ResNet-50	120.2	81.6
	HRNet-W32	118.7	78.0
Our Method (SD + TD)	ResNet-18	98.3	67.4
	ResNet-50	96.3	65.5
	HRNet-W32	**94.7**	**63.9**

Table 5: Human pose estimation performance of different discrepancy distance error on 3DPW dataset.

Discrepancy Distance Error	MPJPE↓	PAMPJPE↓
Naïve l_2 distance	100.3	67.2
Kullback-Leibler divergence [29]	96.5	65.1
Maximum Mean Discrepancy [14]	**94.7**	**63.9**

Table 6: Human pose estimation performance with various number of source datasets on 3DPW and 3DHP dataset.

Source Datasets				Test on 3DPW	
H3.6M	3DHP	HumanEva	SURREAL	MPJPE↓	PAMPJPE↓
√	-	-	-	94.7	63.9
√	√	-	-	94.6	63.8
√	√	√	-	94.4	63.6
√	√	√	√	94.1	63.5

Source Datasets				Test on 3DHP	
3DPW	H3.6M	HumanEva	SURREAL	MPJPE↓	PAMPJPE↓
√	-	-	-	109.7	88.4
√	√	-	-	105.2	84.3
√	√	√	-	104.1	83.6
√	√	√	√	102.7	82.5

method (SD + Multi-UDs) again exhibits a competitive performance against the domain adaptation model and significantly surpassing compared domain adaptation approaches.

5.4. Ablation Study

We study the effectiveness of various components of the proposed method. Unless specified, we train our model on Human3.6M dataset and then validate on the 3DPW dataset.

Backbone Architecture and Loss Functions We first examine the impact of variation in backbone architecture, including ResNet-18, ResNet-50 [17] and HRNet-W32 [60]. Table 4 reports the performance of the naïve baseline model and the proposed model under domain adaptation setting. As shown, HRNet outperforms ResNet-50 and ResNet-18. In addition, the proposed method consistently outperforms the baseline with each backbone, which indicates our method is model-agnostic and can be applied to common architectures. Then we compare the results of various discrepancy distance loss in Table 5. Considering MMD and KL divergence are both frequently used in domain adaptation, it is reasonable to choose either of these two errors instead of l_2 distance. The results indicate that choosing MMD as discrepancy distance error is better.

Variation on Source Dataset This ablation study aims to verify whether increase the number of source datasets can be a practical way to train a better model. We experiment two combinations of source and target dataset, *i.e*. the first (second) combination use Human3.6M (3DPW) as source dataset and 3DPW (3DHP) as target dataset. Then, we gradually used more datasets as source and report the results in Table 6. In the first combination, the target (*i.e*. 3DPW) is an in-the-wild dataset, while the sources mostly comprise indoor controlled environments. The results reveal a key insight, *i.e*. more source datasets do not necessarily enrich the diversity of training data, hence the marginal improvement. In the second combination, we observed up to -7mm MPJPE when more sources are added, as the new sources provide more diverse and complementary information.

Variation on Interventions Here, we examine the effects of mixing few unconstrained datasets as interventions, as well as different types of datasets. In addition to using Human3.6M as source dataset, we also consider action recognition dataset (*i.e*. UCF101 [58], HMDB [28] and Kinetics [25]) and general image dataset (*i.e*. ImageNet [6], PAS-

Table 7: Human pose estimation performance with various number of interventions on 3DPW dataset.

Interventions			MPJPE↓	PAMPJPE↓
UCF101	HMDB	Kinetics		
√	-	-	98.6	67.3
-	√	-	98.1	66.9
-	-	√	97.5	66.4
√	√	√	94.9	64.2
ImageNet	PASCAL	MS COCO		
√	-	-	102.3	69.4
-	√	-	102.8	69.7
-	-	√	101.4	68.9
√	√	√	99.2	67.8

CAL [8] and COCO Panoptic [36]). As shown in Table 7, there are no significant difference in performance when different unconstrained domain datasets are used singly. When all datasets are used as interventions, the MPJPE is improved greatly. We also observe that when using the action recognition datasets, the performance is close to our domain adaptation model (*i.e.* SD + TD). Although the performance with the general image datasets is not as good, it still outperforms the compared domain adaptation models in Table 2.

6. Discussion

In this section, we discuss the reason why increasing source dataset and intervention can improve performance. We first recall the empirical risk minimization setup, where a learning model accesses data from a distribution $P(X,Y)$ and trains a predictor ϕ in a hypothesis space $\mathcal{H}$ to minimize the empirical risk $\hat{R}$:

$$\phi^* = \arg\min_{\phi\in\mathcal{H}} \hat{R}_{P(X,Y)}(\phi) \tag{5}$$

$$\hat{R}_{P(X,Y)}(\phi) = \hat{\mathbb{E}}_{P(X,Y)}[\text{loss}(Y, \phi(X)] \tag{6}$$

Here, we denote by $\hat{\mathbb{E}}_{P(X,Y)}$ the empirical mean computed from a sample drawn from $P(X,Y)$. An out-of-distribution (OOD) generalization means having a small expected risk for a different distribution $P^t(X,Y)$:

$$R^{\text{OOD}}_{P^t(X,Y)}(\phi) = \mathbb{E}_{P^t(X,Y)}[\text{loss}(Y, \phi(X)] \tag{7}$$

Clearly, the gap between $\hat{R}_{P(X,Y)}(\phi)$ and $R^{\text{OOD}}_{P^t(X,Y)}(\phi)$ will depend on how different the test distribution P^t is from the training distribution P. To quantify this difference, we define *domains* as the collection of different circumstances that give rise to the distribution shifts. Domains can be modeled as a causal factorization as they are regarded as interventions on one or several causal variables or mechanism [52]. We could restrict $P^t(X,Y)$ to be the result of a certain set of interventions, *i.e.* $P^t(X,Y) \in \mathbb{P}_\mathcal{G}$ where $\mathbb{P}_\mathcal{G}$ is a set of interventional distributions over a causal graph $\mathcal{G}$. The worst-case out-of-distribution risk then becomes

$$R^{\text{OOD}}_{\mathbb{P}_\mathcal{G}}(\phi) = \max_{P^t\in\mathbb{P}_\mathcal{G}} \mathbb{E}_{P^t(X,Y)}[\text{loss}(Y, \phi(X)] \tag{8}$$

To learn a robust predictor, we should have available a subset of domain distributions $\mathcal{E} \subset \mathbb{P}_\mathcal{G}$ and solve

$$\phi^* = \arg\min_{\phi\in\mathcal{H}} \max_{P^t\in\mathcal{E}} \hat{\mathbb{E}}_{P^t(X,Y)}[\text{loss}(Y, \phi(X)] \tag{9}$$

Learning the model by solving the min-max optimization problem of Eq. 9 is challenging. We utilize several common machine learning techniques to approximate Eq. 9.

The first approach is enriching the distribution of training set. This does not mean obtaining more examples from $P(X,Y)$, but training on a richer dataset. Since this strategy is based on standard empirical risk minimization, it can achieve stronger generalization in practice only if the new training distribution is sufficiently diverse to contain information about other distributions in $\mathbb{P}_\mathcal{G}$. As shown in Section 5.4, the proposed method is able to incorporate more source datasets during training to achieve this.

The second approach is to increase the diversity of interventions. The intuition of the intervention is to encourage a model to learn the underlying invariances or symmetries present in the interventional distributions. As shown in the study of variations on interventions (*cf.* Section 5.4), we specify a set of interventions $\mathcal{E}$ by introducing unconstrained domain datasets to generate counterfactual features to which the model should be robust. Instead of computing the maximum over all distributions in $\mathbb{P}_\mathcal{G}$, we can relax the problem by sampling from the interventional distributions and optimize an expectation over a suitably chosen subset.

7. Conclusion

In this paper, we draw ideas from causality to generatively intervene in the training of robust pose estimation models for cross-domain pose estimation. We consider changing domain as interventions on images under the data-generation process and steer generative model to produce counterfactual features, which help the model learn transferable and causal relations across different domains. With data from single or multiple domains, we demonstrate that our approach can improve performance with unlabeled target domain data, and gain out-of-distribution robustness to unseen data. In principle, the proposed method is applicable to most visual recognition tasks and we plan to verify its effectiveness in other fields in the future.

Acknowledgements This research is supported by the National Key Research and Development Program of China (No.2017YFB1002802).

References

[1] Jayadev Acharya, Arnab Bhattacharyya, Constantinos Daskalakis, and Saravanan Kandasamy. Learning and testing causal models with interventions. In *NeurIPS*, pages 9469–9481, 2018.

[2] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. In *ICLR*, 2021.

[3] Seungryul Baek, Kwang In Kim, and Tae-Kyun Kim. Weakly-supervised domain adaptation via GAN and mesh model for estimating 3D hand poses interacting objects. In *CVPR*, pages 6121–6131, 2020.

[4] Krzysztof Chalupka, Pietro Perona, and Frederick Eberhardt. Visual causal feature learning. In *UAI*, pages 181–190, 2015.

[5] Ching-Hang Chen and Deva Ramanan. 3D human pose estimation = 2D pose estimation + matching. In *CVPR*, pages 7035–7043, 2017.

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009.

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

[8] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2):303–338, 2010.

[9] Tian Gan, Junnan Li, Yongkang Wong, and Mohan Kankanhalli. A multi-sensor framework for personal presentations analytics. *ACM Transactions on Multimedia, Computing Communications and Application*, 15(2):30:1–30:21, 2019.

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

[11] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

[12] Mingming Gong, Kun Zhang, Tongliang Liu, Dacheng Tao, Clark Glymour, and Bernhard Schölkopf. Domain adaptation with conditional transferable components. In *ICML*, pages 2839–2848, 2016.

[13] John C Gower. Generalized procrustes analysis. *Psychometrika*, 40(1):33–51, 1975.

[14] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

[15] Ikhsanul Habibie, Weipeng Xu, Dushyant Mehta, Gerard Pons-Moll, and Christian Theobalt. In the wild human pose estimation using explicit 2D features and intermediate 3D representations. In *CVPR*, pages 10905–10914, 2019.

[16] Negar Hassanpour and Russell Greiner. Counterfactual regression with importance sampling weights. In *IJCAI*, pages 5880–5887, 2019.

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

[18] Christina Heinze-Deml and Nicolai Meinshausen. Conditional variance penalties and domain shift robustness. *Machine Learning*, 110(2):303–348, 2021.

[19] Xianxu Hou, Linlin Shen, Ke Sun, and Guoping Qiu. Deep feature consistent variational autoencoder. In *WACV*, pages 1133–1141, 2017.

[20] Maximilian Ilse, Jakub M Tomczak, and Patrick Forré. Designing data augmentation for simulating interventions. In *ICML*, 2021.

[21] 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.

[22] Fredrik Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In *ICML*, pages 3020–3029, 2016.

[23] Fredrik D Johansson, Nathan Kallus, Uri Shalit, and David Sontag. Learning weighted representations for generalization across designs. *arXiv preprint arXiv:1802.08598*, 2018.

[24] Justin Johnson, Alexandre Alahi, and Fei-Fei Li. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, volume 9906 of *Lecture Notes in Computer Science*, pages 694–711, 2016.

[25] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[26] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

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

[28] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. HMDB: a large video database for human motion recognition. In *ICCV*, pages 2556–2563, 2011.

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

[30] David Lewis. Causation. *The Journal of Philosophy*, 70(17):556–567, 1974.

[31] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. Learning to generalize: Meta-learning for domain generalization. In *AAAI*, volume 32, 2018.

[32] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *ICCV*, pages 5542–5550, 2017.

[33] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *CVPR*, pages 5400–5409, 2018.

[34] Sijin Li and Antoni B Chan. 3D human pose estimation from monocular images with deep convolutional neural network. In *ACCV*, pages 332–347. Springer, 2014.

[35] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, pages 2980–2988, 2017.

[36] 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.

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

[38] David Lopez-Paz, Robert Nishihara, Soumith Chintala, Bernhard Scholkopf, and Léon Bottou. Discovering causal signals in images. In *CVPR*, pages 6979–6987, 2017.

[39] Chengzhi Mao, Amogh Gupta, Augustine Cha, Hao Wang, Junfeng Yang, and Carl Vondrick. Generative interventions for causal learning. In *CVPR*, pages 3947–3956, 2021.

[40] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *ICCV*, pages 2794–2802, 2017.

[41] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3D human pose estimation. In *ICCV*, pages 2640–2649, 2017.

[42] 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.

[43] Jovana Mitrovic, Brian McWilliams, Jacob Walker, Lars Buesing, and Charles Blundell. Representation learning via invariant causal mechanisms. In *ICLR*, 2021.

[44] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *ICML*, pages 10–18, 2013.

[45] 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.

[46] Markus Oberweger and Vincent Lepetit. DeepPrior++: Improving fast and accurate 3d hand pose estimation. In *ICCV Workshops*, pages 585–594, 2017.

[47] Judea Pearl. *Causality*. Cambridge University Press, 2009.

[48] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society*, 75(5):947–1012, 2016.

[49] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

[50] Helge Rhodin, Mathieu Salzmann, and Pascal Fua. Unsupervised geometry-aware representation for 3D human pose estimation. In *ECCV*, volume 11214 of *Lecture Notes in Computer Science*, pages 765–782, 2018.

[51] Axel Sauer and Andreas Geiger. Counterfactual generative networks. In *ICLR*, 2021.

[52] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 2021.

[53] Patrick Schwab, Lorenz Linhardt, and Walter Karlen. Perfect match: A simple method for learning representations for counterfactual inference with neural networks. *arXiv preprint arXiv:1810.00656*, 2018.

[54] Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *ICML*, pages 3076–3085, 2017.

[55] 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.

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

[57] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[58] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[59] Guofei Sun, Yongkang Wong, Zhiyong Cheng, Mohan S. Kankanhalli, Weidong Geng, and Xiangdong Li. Deepdance: Music-to-dance motion choreography with adversarial learning. *IEEE Transactions on Multimedia*, 23:497–509, 2021.

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

[61] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR*, pages 1521–1528, 2011.

[62] Alexander Toshev and Christian Szegedy. DeepPose: Human pose estimation via deep neural networks. In *CVPR*, pages 1653–1660, 2014.

[63] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *CVPR*, pages 7167–7176, 2017.

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

[65] 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.

[66] 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.

[67] 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.

[68] Bingjie Xu, Yongkang Wong, Junnan Li, Qi Zhao, and Mohan Kankanhalli. Learning to detect human-object interactions with knowledge. In *CVPR*, pages 2019–2028, 2019.

[69] Liuyi Yao, Zhixuan Chu, Sheng Li, Yaliang Li, Jing Gao, and Aidong Zhang. A survey on causal inference. *ACM Transactions on Knowledge Discovery from Data*, 15(5):1–46, 2021.

[70] Liuyi Yao, Sheng Li, Yaliang Li, Mengdi Huai, Jing Gao, and Aidong Zhang. Representation learning for treatment effect estimation from observational data. In *NeurIPS*, pages 2638–2648, 2018.

[71] Liuyi Yao, Sheng Li, Yaliang Li, Mengdi Huai, Jing Gao, and Aidong Zhang. ACE: Adaptively similarity-preserved representation learning for individual treatment effect estimation. In *ICDM*, pages 1432–1437, 2019.

[72] Jianfeng Zhang, Xuecheng Nie, and Jiashi Feng. Inference stage optimization for cross-scenario 3D human pose estimation. In *NeurIPS*, 2020.

[73] Xiheng Zhang, Yongkang Wong, Mohan S Kankanhalli, and Weidong Geng. Unsupervised domain adaptation for 3d human pose estimation. In *ACM MM*, pages 926–934, 2019.

[74] Long Zhao, Xi Peng, Yuxiao Chen, Mubbasir Kapadia, and Dimitris N Metaxas. Knowledge as priors: Cross-modal knowledge generalization for datasets without superior knowledge. In *CVPR*, pages 6528–6537, 2020.

[75] Xingyi Zhou, Qixing Huang, Xiao Sun, Xiangyang Xue, and Yichen Wei. Towards 3D human pose estimation in the wild: a weakly-supervised approach. In *ICCV*, pages 398–407, 2017.

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

[77] 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.