Learning Causal Representation for Training Cross-Domain Pose Estimator via Generative Interventions

Xiheng Zhang\textsuperscript{1}, Yongkang Wong\textsuperscript{2}, Xiaofei Wu\textsuperscript{3}, Juwei Lu\textsuperscript{3}, Mohan Kankanhalli\textsuperscript{2}, Xiangdong Li\textsuperscript{1}\ast, Weidong Geng\textsuperscript{1}\ast

\textsuperscript{1}State Key Laboratory of CAD&CG, College of Computer Science and Technology, Zhejiang University
\textsuperscript{2}School of Computing, National University of Singapore
\textsuperscript{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 \cite{9, 59, 68}. In recent years, the deep learning models have achieved tremendous improvement with advance in model architecture \cite{7, 17, 57}, novel loss functions \cite{24, 35}, and availability of quality datasets \cite{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).

The deficiency on cross-domain pose estimation can be attributed to dataset biases \cite{61} or shortcut learning \cite{11}, which means that deep learning models are prone to learn dataset-dependent spurious correlations based on statistical associations \cite{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 \cite{52}. Causal relations reflect the fundamental data-generating mechanism, which tends to be universal and invariant across different domains \cite{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, Johanson 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,
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 enable 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 $PA_i$) and an unexplained random variable $U_i$, i.e.

$$X_i := f_i(PA_i, U_i), (i = 1, ..., n),$$  \hspace{1cm} (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 $\hat{P}^E = (x_i, t_i)_{i=1}^n$ and the counterfactual samples consist of set $\hat{P}^{CF} = (x_i, 1 - t_i)_{i=1}^n$. Here, set $P^E \sim \hat{P}^E$ is the empirical observed distribution and set $P^{CF} \sim \hat{P}^{CF}$ is the empirical counterfactual distribution. As only one potential outcome could be observed, we define the observed outcomes as $y^E(x)$ and the unobserved outcomes as counterfactual outcomes $y^{CF}(x)$.

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.](image-url)
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 \(X\), 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 \(\{x_i, d_i\}\) 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, \(P^\mathcal{T}(X, D) = P(X) \cdot P(D_s | X)\) 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 \(P^\mathcal{T}\).

### 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, locating 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, \(x = M(z) \sim p(x|z)\). The objective is defined as the minimization of the Evidence Lower Bound (ELBO) [27]:

\[
\mathbb{E}_{q(z|x)} [\log p(x|z)] - D_{KL}(q(z|x) \| p(z))
\]
\[
\min_{\theta_E, \theta_M} \mathcal{L} = -\mathbb{E}_{q(z|x)}[\log p(x|z)] + \text{KL}(q(z|x)||p(z))
\]  

where \(\theta_E\) and \(\theta_M\) are the parameters of \(E\) and \(M\), respectively. \(\text{KL}()\) 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 \(D\) cannot reliably distinguish them. Based on the Least Squares GAN [40], the following min-max game is defined:

\[
\begin{align*}
\min_{\theta_g} \max_{\theta_D} \mathcal{L} &= \mathbb{E}_{(u,y) \sim (p(u),p(y))} \|D(g(u, y)) - 1\|^2 \\
&+ \mathbb{E}_{x \sim p(x)} \|D(E(x))\|^2
\end{align*}
\]  

where \(\theta_g\) (\(\theta_D\)) 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:

\[
\begin{align*}
\min_{\theta_f, \theta_h} \mathcal{L} &= \mathbb{E}_{(x,y,u) \sim (p(x),p(y),p(u))} \mathcal{L}_F(h(f(x)), y) \\
&+ \lambda_1 \mathcal{L}_{CF}(h(g(u, y)), y) + \lambda_2 \mathcal{L}_{dist}(f(x), g(u, y))
\end{align*}
\]  

where \(\lambda_1\) and \(\lambda_2\) control the strength of the imbalance penalties. The loss term \(\mathcal{L}_F\) and \(\mathcal{L}_{CF}\) 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 \(\mathcal{L}_{dist}\), we select Maximum Mean Discrepancy [14], KL divergence [29] and naive \(l_2\) 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 \(y_{ik}\). 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>
<thead>
<tr>
<th>Task</th>
<th>Training Set</th>
<th>Test Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conventional Learning</td>
<td>SD</td>
<td>TD</td>
</tr>
<tr>
<td>Domain Adaptation</td>
<td>SD + TD (w/o label)</td>
<td>TD</td>
</tr>
<tr>
<td>Domain Generalization</td>
<td>SD + UD</td>
<td>TD</td>
</tr>
</tbody>
</table>
Table 2: Human pose estimation results. The experiment is conducted on various source→target settings.

<table>
<thead>
<tr>
<th>Learning Category</th>
<th>Methods</th>
<th>H3.6M→3DPW</th>
<th>H3.6M→3DHP</th>
<th>H3.6M→SURREAL</th>
<th>H3.6M→HumanEva</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>MPJPE↓</td>
<td>PAMPJPE↓</td>
<td>MPJPE↓</td>
<td>PAMPJPE↓</td>
</tr>
<tr>
<td>Conventional</td>
<td>Source only</td>
<td>118.7</td>
<td>78.0</td>
<td>121.8</td>
<td>98.5</td>
</tr>
<tr>
<td>Domain Adaptation</td>
<td>DDC [64]</td>
<td>110.4</td>
<td>75.3</td>
<td>115.6</td>
<td>91.5</td>
</tr>
<tr>
<td></td>
<td>DAN [37]</td>
<td>107.5</td>
<td>73.2</td>
<td>109.5</td>
<td>89.2</td>
</tr>
<tr>
<td></td>
<td>DANN [10]</td>
<td>106.3</td>
<td>71.1</td>
<td>107.9</td>
<td>88.0</td>
</tr>
<tr>
<td></td>
<td>ISO [72]</td>
<td>-</td>
<td>70.8</td>
<td>-</td>
<td>75.8</td>
</tr>
<tr>
<td></td>
<td>Our method (SD + TD)</td>
<td>94.7</td>
<td>63.9</td>
<td>99.3</td>
<td>81.5</td>
</tr>
<tr>
<td>Domain Generalization</td>
<td>Wang et al. [67]</td>
<td>109.5</td>
<td>68.3</td>
<td>111.9</td>
<td>89.0</td>
</tr>
<tr>
<td></td>
<td>Our method (SD + UD)</td>
<td>97.5</td>
<td>66.4</td>
<td>102.6</td>
<td>86.4</td>
</tr>
<tr>
<td></td>
<td>Our method (SD + Multi-UDs)</td>
<td>94.9</td>
<td>64.2</td>
<td>101.6</td>
<td>83.7</td>
</tr>
</tbody>
</table>

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 \((-z_{max}, +z_{max})\) 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 minibatch size of 128. The initial learning rate is set to \(1 \times 10^{-3}\) and reduced by a factor of 10 at the 170th epoch. We use \(256 \times 256\) and \(384 \times 288\) as the input size of ResNet and HRNet, respectively. The data augmentation scheme includes random rotation \((-45^\circ, 45^\circ)\), 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.

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

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

<table>
<thead>
<tr>
<th>Method</th>
<th>Backbone</th>
<th>MPJPE ↓</th>
<th>PAMPJPE ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source only</td>
<td>ResNet-18</td>
<td>122.4</td>
<td>83.1</td>
</tr>
<tr>
<td></td>
<td>ResNet-50</td>
<td>120.2</td>
<td>81.6</td>
</tr>
<tr>
<td></td>
<td>HRNet-W32</td>
<td>118.7</td>
<td>78.0</td>
</tr>
<tr>
<td></td>
<td>ResNet-18</td>
<td>98.3</td>
<td>67.4</td>
</tr>
<tr>
<td></td>
<td>ResNet-50</td>
<td>96.3</td>
<td>65.5</td>
</tr>
<tr>
<td></td>
<td>HRNet-W32</td>
<td>94.7</td>
<td>63.9</td>
</tr>
<tr>
<td>Our Method (SD + TD)</td>
<td>ResNet-18</td>
<td>98.3</td>
<td>67.4</td>
</tr>
<tr>
<td></td>
<td>ResNet-50</td>
<td>96.3</td>
<td>65.5</td>
</tr>
<tr>
<td></td>
<td>HRNet-W32</td>
<td>94.7</td>
<td>63.9</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Discrepancy Distance Error</th>
<th>MPJPE ↓</th>
<th>PAMPJPE ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naive $l_2$ distance</td>
<td>100.3</td>
<td>67.2</td>
</tr>
<tr>
<td>Kullback-Leibler divergence [29]</td>
<td>96.5</td>
<td>65.1</td>
</tr>
<tr>
<td>Maximum Mean Discrepancy [14]</td>
<td>94.7</td>
<td>63.9</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Source Datasets</th>
<th>Test on 3DPW</th>
<th>Test on 3DHP</th>
</tr>
</thead>
<tbody>
<tr>
<td>3DPW H3.6M HumanEva SURREAL</td>
<td>MPJPE ↓</td>
<td>PAMPJPE ↓</td>
</tr>
<tr>
<td>√</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>√</td>
<td>√</td>
<td>-</td>
</tr>
<tr>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>3DPW H3.6M HumanEva SURREAL</td>
<td>MPJPE ↓</td>
<td>PAMPJPE ↓</td>
</tr>
<tr>
<td>√</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>√</td>
<td>√</td>
<td>-</td>
</tr>
<tr>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
</tbody>
</table>

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 naive 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.

<table>
<thead>
<tr>
<th>Interventions</th>
<th>ImageNet</th>
<th>PASCAL</th>
<th>MS COCO</th>
</tr>
</thead>
<tbody>
<tr>
<td>UCF101</td>
<td>✓</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>HMDB</td>
<td>-</td>
<td>✓</td>
<td>-</td>
</tr>
<tr>
<td>Kinetics</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>MPJPE↓</td>
<td>98.6</td>
<td>102.3</td>
<td>101.4</td>
</tr>
<tr>
<td>PAMPJPE↓</td>
<td>67.3</td>
<td>69.4</td>
<td>68.9</td>
</tr>
</tbody>
</table>

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 \( P \) and trains a predictor \( \phi \) in a hypothesis space \( \mathcal{H} \) to minimize the empirical risk \( \hat{R} \):

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

Here, we denote by \( \mathbb{E}_{P(X,Y)} \) 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 \( P'(X,Y) \):

\[
R_{P'(X,Y)}^{\text{OOD}}(\phi) = \mathbb{E}_{P'(X,Y)}[\text{loss}(Y, \phi(X))]
\]

Clearly, the gap between \( \hat{R}_{P(X,Y)}(\phi) \) and \( R_{P'(X,Y)}^{\text{OOD}}(\phi) \) will depend on how different the test distribution \( P' \) 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 \( P'(X,Y) \) to be the result of a certain set of interventions, i.e. \( P'(X,Y) \in \mathbb{P}_G \) where \( \mathbb{P}_G \) is a set of interventional distributions over a causal graph \( \mathcal{G} \). The worst-case out-of-distribution risk then becomes

\[
R_{\mathbb{P}_G}^{\text{OOD}}(\phi) = \max_{P' \in \mathbb{P}_G} \mathbb{E}_{P'(X,Y)}[\text{loss}(Y, \phi(X))]
\]

To learn a robust predictor, we should have available a subset of domain distributions \( \mathcal{E} \subset \mathbb{P}_G \) and solve

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

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 \( \mathbb{P}_G \). 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 \( \mathbb{P}_G \), 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


