PoseIRM: Enhance 3D Human Pose Estimation on Unseen Camera Settings via Invariant Risk Minimization

Yanlu Cai[1], Weizhong Zhang[1,*], Yuan Wu[1], Cheng Jin[1,2,*]
[1]Fudan University, Shanghai, China [2]Haina Lab, Shanghai, China
{ylcai20, weizhongzhang, wuyuan, jc}@fudan.edu.cn

Abstract

Camera-parameter-free multi-view pose estimation is an emerging technique for 3D human pose estimation (HPE). They can infer the camera settings implicitly or explicitly to mitigate the depth uncertainty impact, showcasing significant potential in real applications. However, due to the limited camera setting diversity in the available datasets, the inferred camera parameters are always simply hardcoded into the model during training and not adaptable to the input in inference, making the learned models cannot generalize well under unseen camera settings. A natural solution is to artificially synthesize some samples, i.e., 2D-3D pose pairs, under massive new camera settings. Unfortunately, to prevent over-fitting the existing camera setting, the number of synthesized samples for each new camera setting should be comparable with that for the existing one, which multiplies the scale of training and even makes it computationally prohibitive. In this paper, we propose a novel HPE approach under the invariant risk minimization (IRM) paradigm. Precisely, we first synthesize 2D poses from myriad camera settings. We then train our model under the IRM paradigm, which targets at learning a common optimal model across all camera settings and thus enforces the model to automatically learn the camera parameters based on the input data. This allows the model to accurately infer 3D poses on unseen data by training on only a handful of samples from each synthesized setting and thus avoid the unbearable training cost increment. Another appealing feature of our method is that benefited from the capability of IRM in identifying the invariant features, its performance on the seen camera settings is enhanced as well. Comprehensive experiments verify the superiority of our approach.

1. Introduction

3D multi-view Human Pose Estimation (HPE) leverages the camera relationship between multiple viewpoint to mitigate the impact of depth uncertainty. Existing methods primarily rely on camera parameters to construct epipolar geometric constraints between camera viewpoints. Unlike camera-parameter-required methods, camera-parameter-free methods can explicitly or implicitly recover camera parameters during training, thus making them applicable in broader scenarios where camera parameters are unavailable, such as HPE in uncontrolled environments or dynamic HPE with moving cameras. HPCP [21] leverages human pose prior such as bone length to optimize potential camera parameters. Flex [4] models viewpoint-consistent 3D poses by hierarchical skeletal representation. MTF-Transformer [20] leverages temporal information to obtain more accurate camera parameters. Probabilistic Triangulation [11] adopts Monte Carlo sampling to select the camera parameters. These methods have achieved commendable results, closely matching the performance of camera-parameter-free methods under seen camera setting.

However, when generalized to unseen camera settings, camera-parameter-free methods exhibit a great performance drop, in contrast the drop in the camera-parameter-required methods is negligible [20]. We argue that this discrepancy primarily stems from the reason that the inferred camera parameters in the camer-parameter-free methods are always simply hardcoded into the models during training, i.e., they are not adaptable to the input in inference. To be precise, the number of camera viewpoints within existing datasets is rather limited (typically four or eight), whereas previous mainstream models require at least four camera viewpoints as input to address the challenges such as self-occlusion of the human body or inaccuracies in 2D pose estimation, making the number of available camera settings no larger than two. With such extremely limited diversity of camera settings, because of the training imbalance, camera-parameter-free methods tend to memorize the specific camera settings rather than to generalize to arbitrary settings, thus significantly hindering their generalization capabilities across varied camera settings. Our experiments have substantiated this conjecture. Please refer to Tab. 5 for details.

A natural solution to address the above challenge is to ar-

*Corresponding authors

tificially synthesize some samples under massive new camera settings and merge them into the training process. Unfortunately, to prevent over-fitting the existing camera settings, it can be expected that the number of synthesized samples for each new camera setting should be comparable with that for the existing one. This would undoubtedly lead to an unacceptable increment of training time and even make the training computationally prohibitive.

In this paper, to enhance the estimation on unseen camera settings, we propose an effective synthetic data augmented HPE approach under the invariant risk minimization (IRM) paradigm, namely PoseIRM [1]. To be precise, instead of synthesizing images directly, which is almost infeasible due to the high complexity and actually unnecessary, we generate 2D poses from myriad new camera settings by either injecting noise into the 2D poses projected from the 3D ground truth poses or projecting noisy 3D poses into camera planes. A specific proposal is developed to select the camera settings close to the reality. We then train a 3D HPE model under the IRM[1] paradigm, which targets at learning a common optimal model across all camera settings and thus enforces the model to automatically learn the camera parameters based on the input data. This framework allows the model to accurately infer 3D poses on the unseen camera settings by training on only a handful of samples from each synthesized camera setting and thus avoid both the training imbalance issue and the unbearable training cost increment. Another appealing feature of our method is that benefited from the enhanced capability of IRM in feature learning, its performance on the seen camera settings is enhanced as well. Moreover, our learning paradigm can be integrated with general HPE methods flexibly. Comprehensive experiments on both Human3.6M and TotalCapture datasets clearly attest to the superiority of our approach. Specifically, our PoseIRM framework enhances the model's generalization capabilities in unseen camera settings. Additionally, by facilitating the model's adaptation to varied and diverse camera settings, the PoseIRM framework further augments its representational capacity, thereby improving performance in seen camera settings.

Our contributions can be summarized as follows.

- We propose an effective data augmentation paradigm, which is comprised of two 2D pose synthesis algorithms (PSA) and a camera setting selection proposal. It enables us to generate high-quality 2D-3D pose under the camera settings similar to reality.
- We propose an Invariant Risk Minimization (IRM) based approach PoseIRM, which enhances the model's representational capability and generalization ability to adopt to unseen camera settings by promoting consistent performance across all camera settings.
- Experiments demonstrate that our methods achieve state-of-the-art (SOTA) results on the Human3.6M and TotalCapture datasets. Moreover, PoseIRM significantly reduce the performance drop of camera-parameter-free methods in unseen camera settings.

[1]Code and Supplementary materials are available at: https://github.com/DoUntilFalse/PoseIRM

2. Related Work

2.1. Monocular 3D HPE

Given that an infinite number of 3D postures can correspond to a single 2D pose, single-view 3D posture estimation is an ill-posed problem. Monocular methods usually leverage human structure prior [19] to constrain the range of feasible 3D poses, thus reducing the depth uncertainty. SemGCN [31] regards the human skeleton as a graph and uses Graph Convolution Network(GCN) to fuse features between keypoints. Therefore, the feature fusion is performed along the human skeleton, and only the features between adjacent keypoints on the skeleton can be directly fused. SRNet [28] splits keypoints by limbs to enhance feature fusion between keypoints within the same limb, and then recombine them. UGCN [23] proposes a Motion Loss, which introduces explicit temporal motion constraints for multi-frame 3D pose estimation to reduce depth uncertainty. GAST-Net [14] proposes a network structure that combines GCN and self-attention mechanism, allowing features between keypoints to be fused not only along the skeleton graph, but also to adaptively extract connections and fuse between keypoints.

Transformer has demonstrated its powerful capabilities across multiple domains [25, 26]. PoseFormer [33] introduce transformer to 3D HPE. Compared with previous GCN methods, PoseFormer can obtain the attention of different joints and fuse them flexibly rather than fuse them through predefined graph, i.e. human skeleton. PoseFormer also adopt temporal transformer to obtain temporal information. However, due to the computational complexity of the Transformer, which increases quadratically with the number of tokens, PoseFormer struggles to utilize information from a greater number of frames. To address this issue, PoseFormerV2 [32] and MixSTE [29] have optimized the model from two different aspects: low-pass filtering in the frequency domain and the separation of individual keypoints, respectively, making it to accommodate longer sequences. DiffPose [3], Multi-hypothesis DiffPose [7] and D3DP [18] leverage Diffusion to generate feasible 3D poses with 2D input poses as condition.

In summary, monocular methods leverage several human prior to reduce depth uncertainty. However, due to the lack of depth information, the error remains significant.

2.2. Multi-view 3D HPE

2.2.1 Multi-view 3D HPE with Camera Parameter

Camera-parameter-required multi-view methods can obtain camera positions, orientation and the relationships between cameras through camera parameters. Consequently, they can utilize epipolar geometry constraints to recover depth information accurately. For instance, a relative depth relationship in one viewpoint may manifest as a left-right relationship in another viewpoint. This implies that the primary focus of these methods is on how to eliminate the impact of inaccurate 2D keypoints. This is because, without error in the 2D results, the 3D poses can be absolutely accurately determinate using traditional methods (such as Triangulation) when the camera parameters are known. Learnable Triangulation [10] propose a non-parametric yet differentiable triangulation method, which enables gradient propagation back to the 2D posture estimator, thereby optimizing its performance. AdaFuse [30] fuses the heatmaps of different views through epipolar lines to improve the accuracy of 2D keypoint estimation and reduce the impact of outliers. Canonical Fusion [17] first recovers consistent camera-agnostic pose representations to reduce the keypoints error and then fuses them by an efficient triangulation module DLT. Cross-view Fusion [16] proposes the recursive pictorial structure model, which partitions the space into several voxels recursively to obtain more accurate 3D poses by progressively eliminating the influence of outliers. DeepFuse [8] leverage IMU-data to refine the pose results. Epipolar Transformer [6] extends feature fusion from the vicinity of keypoints to the vicinity of the epipolar lines, in order to compensate for misalignments caused by depth uncertainty. In summary, camera-parameter-required multi-view methods can leverage camera parameter to determinate camera positions and relationships, thus utilizing epipolar geometry constraints to recover depth information.

2.2.2 Multi-view 3D HPE without Camera Parameter

Without camera parameter, the camera position, orientation, and relationships need to be regressed by the model. The error in regressing camera parameter makes 3D HPE even more challenging.

HPCP [21] leverages human pose prior such as bone length to optimize potential camera parameters and to predict 3D poses more accurately. FLEX [4] leverage viewpoint-independent skeleton representation(bones and angles) to model a consistent human pose between different views. However, due to error accumulation of hierarchical representation, large errors is presented at terminal keypoints like wrist and ankle. MTF-Transformer [20] leverage transformation matrix regression to force model to learn the camera parameters. Probabilistic Triangulation [11] maintain camera distribution by computing the posterior probability of the camera through Monte Carlo sampling.

However, the number of cameras in the dataset is quite limited (typically 4 or 8), with fixed positions and orientations. Moreover, multi-view methods often take four-view 2d poses as input to mitigate self-occlusion of the human body. This results in the very limited diversity of camera setting, typically restricted to one or two types of camera setting. Under such camera setting with very limited diversity, models tend to memorize the seen settings instead of generalizing to arbitrary camera setting. Thus, we propose a pose synthesis algorithm and a IRM-based pose estimation method, namely PoseIRM, which demonstrate strong generalization capabilities across various camera settings.

2.3. Invariant Risk Minimization

IRM [1] is an emerging machine learning paradigm to enable the learned model generalize on the unseen data with distributional shift. The basic idea is to learn an invariant feature representation on the datasets drawn from multiple environments, in the sense that with this representation one is able to learn a common classifier working well across all these environments, and thus the learned model can be expected to generalize well on the data with unseen distributions. Based on this idea, IRM can be naturally phrased as a bi-leveled optimization problem (please refer to Eqn.(8) for details), however, it is challenging to optimize due to the high computational complexity. Therefore, more practical version of IRM are proposed in the recent studies [13, 34], by relaxing the bi-leveled optimization problem into a regularized minimization task. Promising empirical results are repeatedly reported in the literature [1, 2, 12, 13, 27]. We notice that IRM have not been introduced to the area of pose estimation. One of the main contribution of this paper is that by proposing a novel synthetic data augmented HPE approach under IRM paradigm, we show that IRM can effectively solve the challenges of HPE in generalization on unseen camera setting and training with imbalanced data.

3. Method

In this section, we first introduce our data augmentation paradigm, which is comprised of one camera setting selection proposal and two 2D pose synthesis algorithms. Then we present our pose estimation approach PoseIRM.

3.1. Camera Setting Selection Proposal

In consideration of authenticity, simple randomizing of camera parameters is not reasonable. Therefore, we adhere to the following principles for selecting camera settings:

1. Reasonable roll angle. That is y-axis of the captured images should be nearly parallel to the ground.
2. Reasonable pitch angle. The angle between the camera's optical axis and the ground should not be too large. In

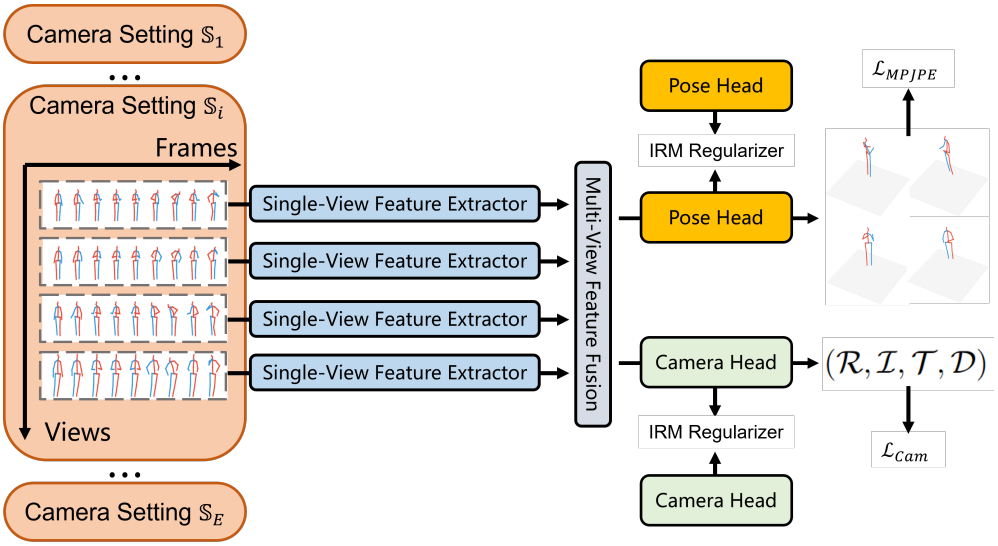

Figure 1. The framework of our PoseIRM. We first generate multiple camera setting $\mathbb{S}_i$, then PoseIRM extract feature from each view and then fuse them to obtain multi-view 3D poses and camera parameters. IRM regularizer is employed to force model to perform consistantly across different camera settings. MPJPE loss and camera loss are employed to supervise the model in regressing 3D poses and camera parameters, respectively.

other words, cameras should capture the subject from the sides rather than from directly above or below.

3. Distinctiveness within the same camera setting: Different viewpoints under the same camera setting should exhibit significant differences.
4. Uniform sampling of camera settings. We uniformly sample the camera settings from all the candidates satisfying the above 3 principles.

Following the above principles, we present a formalized description of the candidate camera setting $\mathbb{S} = \{\mathbb{V}_i\}_{i=1}^{V}$ with V viewpoints:

$$\mathbb{V} = (\mathcal{R}, \mathcal{I}, \mathcal{T}, \mathcal{D}), \tag{1}$$

where $\mathcal{R}$ and $\mathcal{T}$ represent the orientation and translation of the camera, while the $\mathcal{I}$ and $\mathcal{D}$ are the intrinsic parameter matrix and the camera distortion.

To prevent ambiguity, let's first clarify the camera and world coordinate systems. In the camera coordinate system (X_c, Y_c, Z_c, O_c), the positive directions of the x-axis and y-axis correspond to the image's top-to-bottom and left-to-right directions, respectively. The positive direction of the z-axis points from the camera's focal point along the principal axis toward the camera plane. The origin O_c is the focal point of the camera. In the world coordinate system (X_w, Y_w, Z_w, O_w), the x-y plane is parallel to the ground, and z-axis direction is perpendicular to the ground, pointing upwards. The positive x-axis direction and the origin O_w is aligned with that of the given dataset.

The camera's orientation $\mathcal{R}$ is expressed using azimuth, pitch, and roll angles, namely α, β, γ. The azimuth angle α is the angle between X_w and the projection of Z_c onto the $X_w - Y_w$ plane. The pitch angle β is the angle between the Z_c and $X_w - Y_w$ planes. The roll angle γ is the angle between Y_c and $X_w - Y_w$ planes.

Adhering to the aforementioned principles 1, 2, and 4, we can randomly sample the camera orientation as follows:

$$R = Euler_{zxy}(\gamma, \beta, \alpha) \tag{2}$$

where $Euler_{zxy}$ is a function which converts Euler angles to rotation matrix, $\alpha \sim \mathcal{U}[-\pi, \pi), \beta \sim \mathcal{U}[-\pi/6, \pi/6]$, and $\gamma \sim \mathcal{U}[-\pi/N, \pi/N]$ with $N = 36$ and $\mathcal{U}$ being the uniform distribution.

We generate the intrinsic matrix $\mathcal{I}$ using camera focal lengths and centers similar to those in the given dataset, i.e.,

$$\mathcal{I} = \begin{bmatrix} \mathcal{F}_x & 0 & \mathcal{C}_x \\ 0 & \mathcal{F}_y & \mathcal{C}_y \\ 0 & 0 & 1 \end{bmatrix}, \tag{3}$$

where $\mathcal{F}$ and $\mathcal{C}$ are sampled from the two dimensional normal distributions $\mathcal{N}(\bar{F}, \sigma_f)$ and $\mathcal{N}(\bar{C}, \sigma_c)$ separately. Their mean values $\bar{F}$ and $\bar{C}$ are the averaged camera focal length and centre of the given dataset, and σ_f^2 and σ_c^2 are the sample variance. We also sample the distortion $\mathcal{D}$ around the average value of that in the given dataset.

Given that the camera orientation is already determined, the translation $\mathcal{T}$ can be decomposed as the look-at point P_{lookAt} and the distance D between camera and the look-at point. The look-at point P_{lookAt} is sampled around the centroid $\mathcal{M}$ of all the keypoints, i.e., $P_{lookAt} \sim \mathcal{N}(\mathcal{M}, \sigma_p) \subset \mathbb{R}^3$, where the standard deviation σ_p is set to be 0.1.

We let the distance D be slightly greater than the minimum distance D_{min} to keep all keypoints within the screen. That is,

$$D_{min} = \max_{\mathcal{J}^i \in \mathcal{J}} (\max(|\mathcal{F}_x \mathcal{J}_x^i|, |\mathcal{F}_y \mathcal{J}_y^i|) - \mathcal{J}_z^i), \tag{4}$$

$$D \sim \mathcal{U}(\frac{1}{1-\lambda_{scale}} D_{min}, \frac{1}{1-\lambda_{scale}} D_{min} + \lambda_D), \tag{5}$$

where $\mathcal{J}^i$ is the location of the i-th keypoint from the keypoints set $\mathcal{J}$ in the given dataset, which is rotated into camera view, λ_{scale} is the scaling factor to make sure that the entire human body rather than the keypoint is within the screen. For the detailed derivation of the formula, please refer to the supplementary materials. Finally, the camera translation can be expressed as,

$$\mathcal{T} = \begin{bmatrix} 0 \\ 0 \\ D \end{bmatrix} - \mathcal{R} P_{lookAt}. \tag{6}$$

To satisfy principle 3, we let each two views in the same camera setting has a significant difference in azimuth angles, we stipulate that,

$$|\alpha_i - \alpha_j| \geq \epsilon, \forall i, j, i \neq j, \tag{7}$$

where ϵ presents the minimum threshold.

3.2. Pose Synthesis Algorithm

Instead of synthesizing images directly, which is almost infeasible due to the high complexity and actually unnecessary, we aim to generate 2D poses from myriad camera settings. Our basic idea is to project 3D ground-truth poses into synthesized camera viewpoints to build the 2D ground-truth poses (2DGT) in the new viewpoint via classic algorithm. To make the generated poses close to the real data, i.e., the poses detected by 2D pose estimators from real images, we propose the following two distinct Pose Synthesis Algorithms (PSA).

Our first algorithm, known as Distribution-Aware PSA (DAPSA), first collects all 2D detection errors ($\mathcal{E}^{(j)}$) for each joint (e.g., wrists and shoulders) and model these errors using Student's t-distribution. Subsequently, DAPSA samples random biases from these t-distributions and apply them to the 2D ground-truth pose obtained through projection from the respective viewpoint. The synthesized results exhibit a consistent error distribution with the real data.

Our second algorithm, Multi-view Consistent PSA (MVCPSA), first triangulates the 2D detection results into 3D synthesized poses. The discrepancies between these synthesized 3D poses and the 3D ground truth arise from the cumulative effects of deviations in multiple 2D poses. Subsequently, we project these synthesized 3D poses onto a new viewpoint to generate 2D poses with similar detection errors in the real data.

As a result, we obtain samples of (2D, 3D) paired poses in the new viewpoint.

3.3. IRM framework

Let $\mathcal{E} := \{\mathbb{S}_1, \mathbb{S}_2, \cdots, \mathbb{S}_E\}$ be the set of E camera setting. Using the framework of IRM, we can naturally phrase our training problem as follows:

$$\min_{\boldsymbol{\omega} := \{\boldsymbol{v}, \Phi\}} \mathcal{R}(\boldsymbol{\omega}) := \sum_{e \in \mathcal{E}} \mathcal{R}^e(\boldsymbol{v}, \Phi) \tag{8}$$

$$s.t. \boldsymbol{v} \in \arg\min_{\boldsymbol{v}^e} \mathcal{R}^e(\boldsymbol{v}^e, \Phi), \forall e \in \mathcal{E}, \tag{9}$$

where $\mathcal{R}^e$ is the weighted sum of 3D poses error L_{MPJPE} and camera parameters error L_{cam} (See in supplementary materials) of the data from the environment e, $\boldsymbol{v}$ and Φ are the parameters of the regression heads and the backbone, i.e., feature extract and fusion modules. Intuitively, the inner loop enforce the learned features would enable common optimal pose regression heads for all camera setting, while the outer loop optimize the overall accuracy.

Notice that direct solving the above bi-leveled optimization problem is challenging, following the recent studies [1, 13] in IRM, we relaxed it into the following one:

$$\min_{\boldsymbol{\omega}} \tilde{\mathcal{R}}(\boldsymbol{\omega}) := \sum_{e \in \mathcal{E}} \mathcal{R}^e(\boldsymbol{\omega}) + \lambda \mathcal{J}(\boldsymbol{\omega}), \qquad \text{(PoseIRM)}$$

where the regular $\mathcal{J}(\omega)$ takes the form of

$$\mathcal{J}(\omega) := \text{Var}[L_{MPJPE}(\boldsymbol{\omega})] + Var[L_{Cam}(\boldsymbol{\omega})] \tag{10}$$

with Var$[\cdot]$ being the variance promoting similar performance for all environments.

Discussion. Our IRM based approach PoseIRM has the following three appealing features. One is that it treats each environment equally and enforce the model to achieve optimal performance in all environments, thus naturally it naturally addressed the training imbalance issue. To be precise, it allows the model to accurately infer 3D poses by training on only a handful of samples from each synthesized camera setting and thus avoid the unbearable training cost increment. The second is our PoseIRM framework aims at learn a common optimal model across all camera settings, enhancing the model's generalization capabilities in unseen camera settings. Additionally, by facilitating the model's adaptation to varied and diverse camera settings, PoseIRM further augments its representational capability, thereby improving performance in seen camera settings.

3.4. 2D-3D Lifting Network

We select a simple baseline as our 2D-3D lifting network as follows. We adopt PoseFormer [33] as single-view feature extractor, which take 2D poses as input and obtain single-view pose feature from multiple frame. Then we

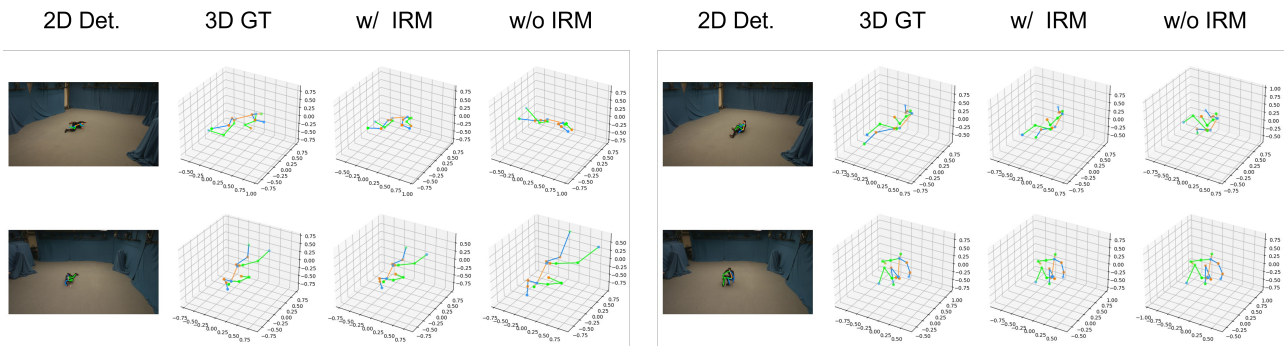

Figure 2. Visualization of the results regressed by the model with/without IRM in the unseen camera setting of TotalCapture dataset.

leverage a two-layer transformer decoder to fuse the four single-view pose feature into multi-view pose feature. After that, two heads are appended to the network. One is for 3D pose regression and the other is to predict camera viewpoint $\mathbb{V}$. Each head adopt three-layer FCN to regress the 3D poses/camera parameters. In consideration of the degrees of freedom, we opt to represent $\mathcal{R}$ using a rotation vector. Specifically, the rotation vector is denoted as $[x*\theta, y*\theta, z*\theta]^T$, where $[x, y, z]^T$ is a unit vector representing the rotation axis and θ signifies the rotation angle. Regarding the camera intrinsic parameter, we employ (fx, fy, cx, cy) to represent the intrinsic matrix $\mathcal{I}$. As for the lens distortion parameters, we utilize $(k1, k2, k3, p1, p2)$ to denote the distortion parameters $\mathcal{D}$.

4. Experiment

We evaluate the performance of our method on the Human3.6M [9] and TotalCapture [22] datasets. We also provide comprehensive ablation studies to demonstrate the effectiveness of our PSA and our IRM architecture. Additionally, we present a camera parameter regression experiment to verify our hypothesis.

4.1. Experimental Settings

Datasets. We evaluate our PoseIRM on both Human3.6M and TotalCapture datasets. Human3.6M dataset is the most widely used 3D Human Pose Estimation dataset, containing over 3 million frames of images with 11 human subjects performing 15 different types of actions captured from four different camera viewpoints. The 3D poses are captured by a motion capture system. Following the settings of previous methods [15, 20, 33], we use 17 keypoints to represent the human pose and use subjects S1, S5, S6, S7, and S8 for training, and subjects S9 and S11 for test. The TotalCapture dataset utilizes 8 completely synchronized cameras to collect four types of actions (Rom, Acting, Walking, and Freestyle) from five subjects, with each action repeated three times, totaling approximately one million frames. We fully follow the experimental settings of the MTF-Transformer [20], where cameras 1, 3, 5, and 7 are used in both the training and test sets, while cameras 2, 4, 6, and 8 appear only in the test set.

Evaluation Metrics. Mean Per Joint Position Error (MPJPE) and Procrustes-aligned MPJPE (P-MPJPE) [24] are used as the evaluation metrics.

Data Augmentation. We generate 7500 different camera settings for Human3.6M datasets and 8100 different camera settings for TotalCapture. The 3D poses of each action clip is processed using a different camera setting for 100 epochs. **To verify our method can infer the camera setting parameter implicitly in inference instead of memorize them, the unseen camera settings in test time are excluded from the generated settings for training.**

4.2. Quantitative Analysis

We report the comparison between our method and other methods with/without camera parameter. As shown in Tab. 1, our method has achieved state-of-the-art (SOTA) performance on Human3.6M. This underscores the effectiveness of our IRM framework in enhancing the model's generalization across various camera settings. It forces the model to extract superior features rather than memorizing a specific camera setting, thereby improving the model's representational capacity. We also report the performance of the methods with/without camera parameter on TotalCapture in Tab. 2. The experimental results indicate that our method exhibits a significant advantage in both seen and unseen camera settings. This further underscores the effectiveness of our IRM framework in enhancing the model's representational capacity. Moreover, our method demonstrates a lower performance drop in unseen camera setting compared to seen camera setting (3.8 v.s. 11.9), approaching the performance drop of camera-parameter-required methods (3.8 v.s. 1.9). Taking into account that the camera-parameter-required methods obtain the camera parameters for new viewpoints, essentially granting it prior exposure to

Method	Input Setting	Dir.	Disc.	Eat.	Greet	Phone	Photo	Pose	Purch.	Sit.	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
Monocular methods																	
SRNet	(CPN, T = 243)	46.6	47.1	43.9	41.6	45.8	49.6	46.5	40.0	53.4	61.1	46.1	42.6	43.1	31.5	32.6	44.8
UGCN	(CPN, T = 96)	40.2	42.5	42.6	41.1	46.7	56.7	41.4	42.3	56.2	60.4	46.3	42.2	46.2	31.7	31.0	44.5
PoseFormer	(CPN, T = 81)	41.5	44.8	39.8	42.5	46.5	51.6	42.1	42.0	53.3	60.7	45.5	43.3	46.1	31.8	32.2	44.3
MHFormer	(CPN, T = 351)	39.2	43.1	40.1	40.9	44.9	51.2	40.6	41.3	53.5	60.3	43.7	41.1	43.8	29.8	30.6	43.0
Occlusion-aware Network	(CPN, T = 128)	38.3	41.3	46.1	40.1	41.6	51.9	41.8	40.9	51.5	58.4	42.2	44.6	41.7	33.7	30.1	42.9
Multi-view Methods with Camera Parameter																	
DeepFuse	(*, T = 1)	26.8	32.0	25.6	52.1	33.3	42.3	25.8	25.9	40.5	76.6	39.1	54.5	35.9	25.1	24.2	37.5
Canonical Fusion	(*, T = 1)	27.3	32.1	25.0	26.5	29.3	35.4	28.8	31.6	36.4	31.7	31.2	29.9	26.9	33.7	30.4	30.2
Epipolar transformers	(*, T = 1)	25.7	27.7	23.7	24.8	26.9	31.4	24.9	26.5	28.8	31.7	28.2	26.4	23.6	28.3	23.5	26.9
Crossview Fusion	(*, T = 1)	24.0	26.7	23.2	24.3	24.8	22.8	24.1	28.6	32.1	26.9	31.0	25.6	25.0	28.0	24.4	26.2
TransFusion	(*, T = 1)	24.4	26.4	23.4	21.1	25.2	23.2	24.7	33.8	29.8	26.4	26.8	24.2	23.2	26.1	23.3	25.8
Learnable Triangulation	(*, T = 1)	19.9	20.0	18.9	18.5	20.5	19.4	18.4	22.1	22.5	28.7	21.2	20.8	19.7	22.1	20.2	20.8
AdaFuse	(*, T = 1)	17.8	19.5	17.6	20.7	19.3	16.8	18.9	20.2	25.7	20.1	19.2	20.5	17.2	20.5	17.3	19.5
MvP	(*, T = 1)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	18.6
MTF-Transformer+	(CPN, T = 27)	23.4	25.2	23.1	24.4	27.4	28.5	22.8	25.2	28.7	36.2	25.9	23.6	26.6	22.6	22.7	25.8
Multi-view Methods without Camera Parameter																	
FLEX	(ResNet152, T = 27)	23.1	28.8	26.8	28.1	31.6	37.1	25.7	31.4	36.5	39.6	35.0	29.5	35.6	26.8	26.4	30.9
FLEX	(CPN, T = 27)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	31.7
Probabilistic Triangulation	(*, T = 1)	24.0	25.4	26.6	30.4	32.1	20.1	20.5	36.5	40.1	29.5	27.4	27.6	20.8	24.1	22.0	27.8
MTF-Transformer	(CPN, T = 27)	23.1	25.4	24.7	24.5	27.9	28.3	23.9	24.6	30.7	35.7	25.8	24.2	28.4	22.8	23.1	26.2
Ours	(CPN, T = 27)	22.2	24.6	22.9	23.2	26.0	27.0	22.2	23.7	29.3	33.6	25.6	22.8	25.8	22.5	22.5	25.1

Table 1. MPJPE on Human3.6M for both camera-parameter-required and camera-parameter-free methods. CPN is adopted as the 2D pose estimator. * means it's an image-to-3d method and no 2D pose estimator is used. T is the number of input frames.

these unseen camera settings, it achieves a very low performance drop. However, our method without camera parameter can also exhibit a comparatively low performance drop, indicating that it has already achieved excellent generalization across different camera settings.

4.3. Generalization

Given the constraint that the Human3.6M dataset only encompasses four camera viewpoints and that prevalent multi-view methods typically accept four viewpoints as input, it becomes challenging to solely rely on Human3.6M for validating the model's generalizability across different camera setting as there are only one setting. To address this, we treat Human3.6M dataset as one camera setting, viewpoint 1,3,5,7 and viewpoint 2,4,6,8 of the TotalCapture dataset as two other camera settings. Thus, we train our model on the Human3.6M dataset and subsequently freeze the model but only fine-tune the pose regression head on viewpoints 1, 3, 5, and 7 of the TotalCapture dataset to adapt to the distinct data initialization of TotalCapture. For instance, the axis pointing vertically upwards from the ground in the TotalCapture dataset differs from that in the Human3.6M dataset. Moreover, to align the skeletal structure of the TotalCapture dataset with that of Human3.6M, we introduced an additional virtual 'Thorax' point to the TotalCapture skeleton. This point is positioned at the midpoint between the Neck and Spine. After fine-tuning the regression head on the TotalCapture dataset, we evaluate the model's generalization capabilities on viewpoints 2, 4, 6, and 8 of TotalCapture. As shown in Tab. 2 with the § mark, our method demonstrates excellent generalizability, which strongly indicates its capability to extract valuable features, enabling the model to swiftly adapt to new camera settings.

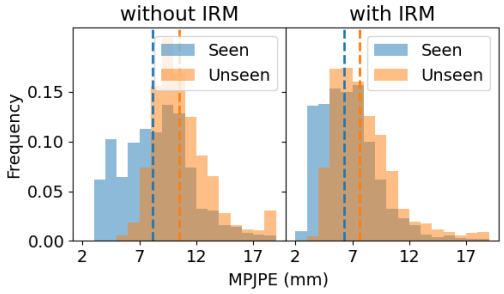

Figure 3. Histogram of the error distribution, which includes the model's performance in Seen/Unseen camera settings with and without Invariant Risk Minimization framework. To eliminate the effects of cross-subject and cross-action variations, we give the results of all samples with seen subjects and seen actions (i.e., excluding freestyle).

4.4. Ablation Study

We performed all our ablation study in the Total Capture dataset. We conducted a comparison of the two PSA methods on the unseen camera setting, and the results are presented in Tab. 3. We can observe that both of our PSA methods have a substantial impact (16.6 and 20.7 vs 27.2). Furthermore, MVCPSA outperforms DAPSA (16.6 vs 20.7). This supports our hypothesis that the errors of 2D pose estimators between different viewpoints are not independently and identically distributed but rather exhibit correlations.

We also conducted ablation experiments on the IRM framework, and the results are presented in Tab. 4 and Fig. 3. In Table 4, we provide the performance of our method with and without IRM in both Seen and Unseen camera settings. The results show that using IRM improves the model's performance in both Seen and Unseen settings. However, we believe that the reasons behind the improvements in these two settings are different. We at-

Method	Seen Cameras(1,3,5,7)							Unseen Cameras(2,4,6,8)							Mean
	Seen Subjects(S1,S2,S3)			Unseen Subjects(S4,S5)			Mean	Seen Subjects(S1,S2)			Unseen Subjects(S4,S5)			Mean	
	W2	FS3	A3	W2	FS3	A3		W2	FS3	A3	W2	FS3	A3		
Crossview Fusion(*)†	19.0	28.0	21.0	32.0	54.0	33.0	29.0	-	-	-	-	-	-	-	-
Canonical Fusion(*)†	10.6	30.4	16.3	27.0	65.0	34.2	27.5	22.4	47.1	27.8	39.1	75.7	43.1	38.2	32.9
MTF-Transformer+(Res101)†	10.7	26.5	16.7	27.4	49.4	34.1	25.1	13.9	29.2	18.1	29.2	49.5	35.6	27.0	26.1
FLEX(Res101)	33.2	81.0	34.2	38.3	123.8	59.5	49.4	109.3	152.1	105.3	114.3	175.5	122.5	125.4	87.4
MTF-Transformer(Res101)§	13.2	31.8	17.1	26.6	51.8	33.0	30.5	34.0	68.0	39.6	47.0	90.7	55.1	56.1	43.3
MTF-Transformer(Res101)	9.3	26.5	14.5	26.7	53.1	33.8	24.7	23.7	40.3	27.4	37.0	61.8	42.9	36.6	30.7
Ours(Res101)§	6.65	19.9	9.47	16.6	29.6	20.1	17.8	10.2	38.3	11.6	16.7	41.9	21.1	23.6	20.7
Ours(Res101)	**5.13**	**14.0**	**7.47**	**14.3**	**25.5**	**18.0**	**14.7**	**6.8**	**24.3**	**9.1**	**14.0**	**35.3**	**19.0**	**18.5**	**16.6**

Table 2. MPJPE on TotalCapture [9] dataset. We adopt Res101 [5] as the 2D pose estimator for fair comparison. † marks camera-parameter-required methods. § presents that we only train the regression head on TotalCapture.

Method	MPJPE	PMPJPE
W/o PSA	27.2	22.0
DAPSA	18.7	13.2
MVCPSA	**16.6**	**12.6**

Table 3. Ablation study of PSA methods. Best in **Bold**.

tribute the enhancement in the Unseen camera setting primarily to IRM improving the model's generalization capabilities, while the improvement in the Seen camera setting is attributed to IRM enhancing the model's feature extraction capabilities. As demonstrated in Tab. 4, there is a gradual enhancement in the model's performance under unseen camera settings as the Penalty term is incrementally increased. This trend underscores the efficacy of Invariant Risk Minimization (IRM) in augmenting the model's generalization capabilities to novel camera configurations. The model achieves its best performance in seen camera settings when the penalty term is optimally balanced, as this facilitates the enhancement of the model's feature representation by forcing model to adapt to more complex and diverse camera settings. However, when the penalty term becomes sufficiently large, the model, in striving for strong consistency across different camera settings, slightly sacrifices performance in settings where it previously excelled, in order to align with the unseen camera settings.

4.5. Camera Parameter Regression

To validate our hypothesis that a limited numver of camera settings leads models to memorize specific camera settings rather than generalize to arbitrary settings, we conduct a camera parameter regression experiment. We conducted comparative analyses with the single-view method PoseFormer and the multi-view method MTF-Transformer. Their method are augmented with an additional camera parameter regression head same as ours. Subsequently, we train the entire model on the TotalCapture dataset. The performance of these methods in regressing camera parameters was then evaluated under both seen and unseen camera settings, with the results detailed in Tab. 5. The results indicate that in the absence of PSA, i.e., under limited camera settings, there is a consistently higher regression error across the models. Notably, under unseen camera settings, the error without PSA approaches half of the disparity observed between seen and unseen camera settings, which is 30°, 7°, and 2m. It implies that their methods only memorize the seen camera settings.

Method	Seen camera setting		Unseen camera setting	
	MPJPE	PMPJPE	MPJPE	PMPJPE
W/o IRM	16.8	13.2	22.4	16.8
IRM(P=1)	**13.2**	**10.0**	20.2	15.2
IRM(P=100)	14.7	11.2	**18.4**	**13.8**

Table 4. Ablation study of IRM framework. Best in **Bold**.

Methods	Seen Camera setting			Unseen Camera setting		
	Axis	Angle	Trans.	Axis	Angle	Trans.
PoseFormer	20°	11°	1.1m	28°	19°	1.5m
MTF-Transformer	6°	5°	0.3m	20°	15°	0.8m
Ours without PSA	4°	2°	0.2m	15°	10°	0.7m
Ours	**1°**	**1°**	**0.04m**	**2°**	**3°**	**0.1m**

Table 5. The results of camera parameter regression. Best in **Bold**. Axis, Angle and Trans. represents the average error of rotation axis, rotation angle and translation.

5. Conclusions

In this paper, to enhance the estimation on unseen camera settings, we propose an effective synthetic data augmented 3D human pose estimation approach under the invariant risk minimization paradigm. Our method PoseIRM allows the model to accurately infer 3D poses on unseen data by training on only a handful of samples from each synthesized setting. Benefited from the strong feature learning capability of IRM, PoseIRM further improves its performance in seen camera settings. Extensive experimental results demonstrate the superiority of our method over the baselines.

6. Acknowledgments

This work was supported by National Natural Science Fund of China (62176064). The computations in this research were performed on the CFFF platform of Fudan University.

References

[1] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. 2, 3, 5

[2] Shiyu Chang, Yang Zhang, Mo Yu, and Tommi Jaakkola. Invariant rationalization. In *International Conference on Machine Learning*, pages 1448–1458. PMLR, 2020. 3

[3] Jia Gong, Lin Geng Foo, Zhipeng Fan, Qiuhong Ke, Hossein Rahmani, and Jun Liu. Diffpose: Toward more reliable 3d pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13041–13051, 2023. 2

[4] Brian Gordon, Sigal Raab, Guy Azov, Raja Giryes, and Daniel Cohen-Or. Flex: Extrinsic parameters-free multi-view 3d human motion reconstruction. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIII*, pages 176–196. Springer, 2022. 1, 3

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. 8

[6] Yihui He, Rui Yan, Katerina Fragkiadaki, and Shoou-I Yu. Epipolar transformers. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 7779–7788, 2020. 3

[7] Karl Holmquist and Bastian Wandt. Diffpose: Multi-hypothesis human pose estimation using diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15977–15987, 2023. 2

[8] Fuyang Huang, Ailing Zeng, Minhao Liu, Qiuxia Lai, and Qiang Xu. Deepfuse: An imu-aware network for real-time 3d human pose estimation from multi-view image. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 429–438, 2020. 3

[9] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36:1325–1339, 2013. 6, 8

[10] Karim Iskakov, Egor Burkov, Victor Lempitsky, and Yury Malkov. Learnable triangulation of human pose. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7718–7727, 2019. 3

[11] Boyuan Jiang, Lei Hu, and Shihong Xia. Probabilistic triangulation for uncalibrated multi-view 3d human pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14850–14860, 2023. 1, 3

[12] Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Domain extrapolation via regret minimization. *arXiv preprint arXiv:2006.03908*, 2020. 3

[13] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*, pages 5815–5826. PMLR, 2021. 3, 5

[14] Junfa Liu, Juan Rojas, Yihui Li, Zhijun Liang, Yisheng Guan, Ning Xi, and Haifei Zhu. A graph attention spatio-temporal convolutional network for 3d human pose estimation in video. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3374–3380. IEEE, 2021. 2

[15] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7753–7762, 2019. 6

[16] Haibo Qiu, Chunyu Wang, Jingdong Wang, Naiyan Wang, and Wenjun Zeng. Cross view fusion for 3d human pose estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4342–4351, 2019. 3

[17] Edoardo Remelli, Shangchen Han, Sina Honari, Pascal Fua, and Robert Wang. Lightweight multi-view 3d pose estimation through camera-disentangled representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6040–6049, 2020. 3

[18] Wenkang Shan, Zhenhua Liu, Xinfeng Zhang, Zhao Wang, Kai Han, Shanshe Wang, Siwei Ma, and Wen Gao. Diffusion-based 3d human pose estimation with multi-hypothesis aggregation. *arXiv preprint arXiv:2303.11579*, 2023. 2

[19] Mingyi Shi, Kfir Aberman, Andreas Aristidou, Taku Komura, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. Motionet: 3d human motion reconstruction from monocular video with skeleton consistency. *ACM Transactions on Graphics (TOG)*, 40:1–15, 2020. 2

[20] Hui Shuai, Lele Wu, and Qingshan Liu. Adaptive multi-view and temporal fusing transformer for 3d human pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 1, 3, 6

[21] Kosuke Takahashi, Dan Mikami, Mariko Isogawa, and Hideaki Kimata. Human pose as calibration pattern: 3d human pose estimation with multiple unsynchronized and uncalibrated cameras. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1856–18567, 2018. 1, 3

[22] Matt Trumble, Andrew Gilbert, Charles Malleson, Adrian Hilton, and John Collomosse. Total capture: 3d human pose estimation fusing video and inertial sensors. In *2017 British Machine Vision Conference (BMVC)*, 2017. 6

[23] Jingbo Wang, Sijie Yan, Yuanjun Xiong, and Dahua Lin. Motion guided 3d pose estimation from videos. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, pages 764–780. Springer, 2020. 2

[24] Jinbao Wang, Shujie Tan, Xiantong Zhen, Shuo Xu, Feng Zheng, Zhenyu He, and Ling Shao. Deep 3d human pose estimation: A review. *Computer Vision and Image Understanding*, 210:103225, 2021. 6

[25] Yibin Wang, Yuchao Feng, Jie Wu, Honghui Xu, and Jianwei Zheng. Ca-gan: Object placement via coalescing attention based generative adversarial network. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pages 2375–2380. IEEE, 2023. 2

[26] Yibin Wang, Haixia Long, Qianwei Zhou, Tao Bo, and Jianwei Zheng. Plsnet: Position-aware gcn-based autism spectrum disorder diagnosis via fc learning and rois sifting. *Computers in Biology and Medicine*, 163:107184, 2023. 2

[27] Chuanlong Xie, Fei Chen, Yue Liu, and Zhenguo Li. Risk variance penalization: From distributional robustness to causality. *arXiv e-prints*, pages arXiv–2006, 2020. 3

[28] Ailing Zeng, Xiao Sun, Fuyang Huang, Minhao Liu, Qiang Xu, and Stephen Lin. Srnet: Improving generalization in 3d human pose estimation with a split-and-recombine approach. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 507–523. Springer, 2020. 2

[29] Jinlu Zhang, Zhigang Tu, Jianyu Yang, Yujin Chen, and Junsong Yuan. Mixste: Seq2seq mixed spatio-temporal encoder for 3d human pose estimation in video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13232–13242, 2022. 2

[30] Zhe Zhang, Chunyu Wang, Weichao Qiu, Wenhu Qin, and Wenjun Zeng. Adafuse: Adaptive multiview fusion for accurate human pose estimation in the wild. *International Journal of Computer Vision*, 129:703–718, 2021. 3

[31] Long Zhao, Xi Peng, Yu Tian, Mubbasir Kapadia, and Dimitris N Metaxas. Semantic graph convolutional networks for 3d human pose regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3425–3435, 2019. 2

[32] Qitao Zhao, Ce Zheng, Mengyuan Liu, Pichao Wang, and Chen Chen. Poseformerv2: Exploring frequency domain for efficient and robust 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8877–8886, 2023. 2

[33] Ce Zheng, Sijie Zhu, Matias Mendieta, Taojiannan Yang, Chen Chen, and Zhengming Ding. 3d human pose estimation with spatial and temporal transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11656–11665, 2021. 2, 5, 6

[34] Xiao Zhou, Yong Lin, Weizhong Zhang, and Tong Zhang. Sparse invariant risk minimization. In *International Conference on Machine Learning*, pages 27222–27244. PMLR, 2022. 3