CEE-Net: Complementary End-to-End Network for 3D Human Pose Generation and Estimation

Haolun Li, Chi-Man Pun*

University of Macau, Macau, China
lhl219319@gmail.com, cmpun@umac.mo

Abstract

The limited number of actors and actions in existing datasets make 3D pose estimators tend to overfit, which can be seen from the performance degradation of the algorithm on cross-datasets, especially for rare and complex poses. Although previous data augmentation works have increased the diversity of the training set, the changes in camera viewpoint and position play a dominant role in improving the accuracy of the estimator, while the generated 3D poses are limited and still heavily rely on the source dataset. In addition, these works do not consider the adaptability of the pose estimator to generated data, and complex poses will cause training collapse. In this paper, we propose the CEE-Net, a Complementary End-to-End Network for 3D human pose generation and estimation. The generator extremely expands the distribution of each joint-angle in the existing dataset and limits them to a reasonable range. By learning the correlations within and between the torso and limbs, the estimator can combine different body-parts more effectively and weaken the influence of specific joint-angle changes on the global pose, improving the generalization ability. Extensive ablation studies show that our pose generator greatly strengthens the joint-angle distribution, and our pose estimator can utilize these poses positively. Compared with the state-of-the-art methods, our method can achieve much better performance on various cross-datasets, rare and complex poses.

Introduction

3D Human Pose Estimation (HPE) is the foundation of many research fields such as action understanding (Yan, Xiong, and Lin 2018; Duan et al. 2022), virtual humans (Guzov et al. 2021), and self-driving (Xu et al. 2021; Bouazizi et al. 2022). Although some latest algorithms (Zeng et al. 2021; Zheng et al. 2021; Li et al. 2022b) have achieved high accuracy on public datasets, their performance on cross-datasets has dropped significantly. In order to maintain stability in reliability-critical applications, some works (Chen et al. 2020; Zeng et al. 2020; Gong et al. 2022; Zhang, Nie, and Feng 2020) improve robustness by analyzing the pose estimator individually. In addition, the insufficient training set is also the main reason for poor generalization ability: **(1)** limited actions restrict the diversity of relative 3D human poses, **(2)** limited actors restrict the range of bone lengths, **(3)** limited cameras restrict the diversity of viewpoints, and **(4)** uneven data distribution makes the network easier to learn common poses while ignoring the rare poses.

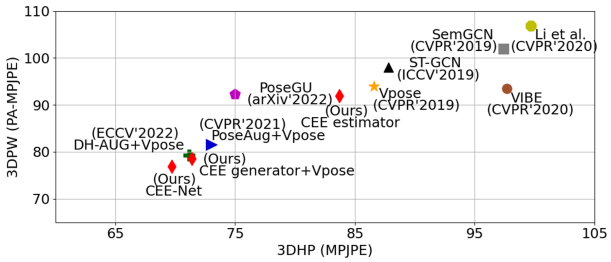

Figure 1: Comparison of the performance on cross-datasets 3DHP and 3DPW between the CEE-Net and state-of-the-art generalization frameworks for 3D human pose estimation. A lower MPJPE and PA-MPJPE indicate better performance.

Large-scale training with sufficient data is the most stable way to improve the generalization of algorithms, but it is limited by collection costs. Fortunately, data generation can be an excellent substitute. When expanding human 3D pose datasets, some papers recombine the joints of different poses in the existing datasets or pre-define a skeleton and rotate joint-angles to generate a large amount of new data (Li et al. 2020; Guan et al. 2022). However, these methods of randomly generating data are easy to generate invalid poses that duplicate the original dataset. In contrast, (Gong, Zhang, and Feng 2021) proposes using Generative Adversarial Networks (GAN) to achieve data augmentation for the first time, which can make the pose generator and the estimator train together. Based on PoseAug (Gong, Zhang, and Feng 2021), (Gholami et al. 2022; Lin, Liang, and Deng 2022) propose various continuous pose generation algorithms, which can be used to train video-based 3D pose estimators. However, through ablation experiments, it is found that the generalization ability of the improved algorithms is almost derived from the new camera viewpoint through Rigid Transformation (RT) and human Bone Length (BL), while the change of Bone Angle (BA) hardly improves the performance. Take PoseAug as an example, and the results are shown in Figure 2. There are three main reasons for

*Corresponding author

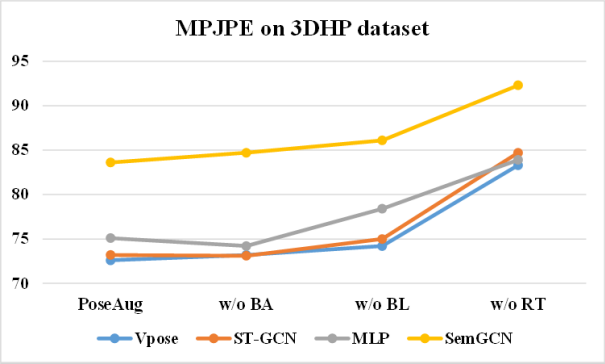

Figure 2: The ablation analysis of four estimators (Cai et al. 2019; Martinez et al. 2017; Pavllo et al. 2019; Zhao et al. 2019) on the 3DHP dataset after BA, BL, and RT are separately removed from the complete PoseAug generator.

this phenomenon: **(1)** In the pose generation part, the diversity of positive samples given to the discriminator is limited, resulting in generated poses being similar to the original. **(2)** Although the Feedback loss (Gong, Zhang, and Feng 2021) seems very reasonable, the existing estimators will lead to it reducing the complexity of generating poses. **(3)** Previous methods do not consider the adaptability of the pose estimator to generated data, and complex poses will cause training collapse.

Based on the above observations, this paper focuses on generating more diverse 3D relative poses and making the pose estimator effectively train the new poses. Then, a novel complementary end-to-end 3D human pose generation and estimation network, named CEE-Net, is proposed. In the pose generation part, we propose an adaptive interpolation sub-net, which increases the diversity of positive samples in the discriminator by using learnable skeleton interpolation between different action poses, and extremely expands the distribution range of each joint-angle of the generated poses. At the same time, to prevent generating unreasonable poses, we design a joint-angle loss, which can control the new joint-angle within the limited range without post-processing selection. In the pose estimation part, we propose a new MLP-based body-parts grouping strategy, which weakens the influence of specific joint-angle changes on global pose estimation. It has strong robustness to complex poses composed of different body-parts and is beneficial to generators. Experiments reflect that our end-to-end pose generator and estimator are mutually promoting. As shown in Figure 1, the enhanced pose estimator performs best in various cross-domain test sets.

In summary, the main contributions of this paper are as follows:

- To the best of our knowledge, we are the first to propose a novel Complementary End-to-End Network (CEE-Net), which efficiently integrates the pose generator and estimator in a co-training framework.
- The adaptive interpolation discriminator in our generator can extremely improve the distribution of each joint-angle, providing the estimator with diverse training data.
- The body-parts grouping strategy in our estimator can effectively utilize these new poses, promoting the generator to output more rare poses.
- Experiments show that our CEE-Net outperforms other state-of-the-art methods in both cross-domain quantitative tests and complex pose qualitative tests.

Related Work

2D-to-3D Pose Estimation

Recovering the 3D human pose can be divided into end-to-end and multi-stage two mainstream frameworks. The end-to-end methods are devoted to estimating the 3D pose from the image directly, but it requires learning overly complex features. The multi-stage-based methods reduce the difficulty by decoupling, which estimate each 2D joint coordinates first and then regress them to the 3D pose. Since the 2D pose estimation is relatively mature, most recent studies focus on improving 2D-to-3D pose estimation. As a simple and effective network structure, the Multi-Layer Perceptron (MLP) has been proven suitable for dealing with low-dimensional tasks in many MLP-based pose estimation algorithms (Martinez et al. 2017; Pavllo et al. 2019; Zeng et al. 2020; Chen et al. 2021; Shan et al. 2021; Wehrbein et al. 2021; Li et al. 2022a). The Graph Convolutional Network (GCN) is composed of nodes and edges, which can correspond to human joints and skeletons, so the GCN-based methods (Zhao et al. 2019; Liu et al. 2020; Xu and Takano 2021; Zou and Tang 2021; Zeng et al. 2021) have skeleton prior information. In addition, since Transformer (Vaswani et al. 2017) is developing rapidly in computer vision, Transformer-based 3D pose estimators also demonstrate strong performance (Zheng et al. 2021; Li et al. 2022b; Chen et al. 2022). Although the accuracy of these SOTA models on many public datasets is getting closer to the theoretical optimum, the accuracy often significantly decreases when deployed into new environments where actors, camera viewpoints, and human actions differ from the training set.

Generalization in 3D Pose Estimator

Similar to the long-tail distribution problem in visual recognition, it is a challenge for the 3D pose estimator when human poses are rare or unseen in the training set. Therefore, some algorithms improve the generalization ability by changing the training strategy. For example, (Chen et al. 2020) enable the network to automatically select the appropriate architecture to estimate heterogeneous human body-parts. (Zeng et al. 2020) split the 2D human pose into local areas and processes them in a separate network branch, weakening the connections between joints. (Zhang, Nie, and Feng 2020) propose the inference stage optimization to mine distributional knowledge about the target scenario. (Guan et al. 2022) propose an unbiased learning method to uniformly learn poses of different actions in a dataset. (Gong et al. 2022) can rely on no dataset through self-supervised learning. Although these methods can improve robustness, it

is difficult to compensate for the shortcomings of the training set, and self-supervised training can not show competitive performance on individual datasets.

Pose Augmentation

The generalization ability of 2D-to-3D pose estimation has been a hot topic in recent years. Due to the limitations of poses in existing datasets and the difficulty of 3D pose collection, using generators to get more 2D-3D pose pairs is a good data augmentation method. (Li et al. 2020) propose using body-parts crossover and mutation operations to generate new 3D poses, but the crossover only between the source dataset, and the fixed range in the mutation is not sufficient. (Gong et al. 2022) use reinforcement learning to control joint-angle rotation without any original dataset, but the performance of self-supervised is limited. (Lin, Liang, and Deng 2022) and (Gong, Zhang, and Feng 2021) investigate the Generative Adversarial Network (GAN) to achieve pose augmentation. Although the performance on the cross-datasets is well improved, their ablation experiments show that the improvement is almost due to the generated camera viewpoints and human positions. (Gholami et al. 2022) uses 2D pose labels of the target domain to generate 2D-3D pose pairs close to their distribution, but the generated dataset is too specialized and requires high quality of the target domain dataset. In general, the 3D poses generated by these works have limited diversity, still heavy reliance on the source dataset, and do not consider the adaptability of the pose estimator to generated data, and complex poses will cause training collapse.

Method

Overview

To address the issue of cross-datasets and rare poses, we propose a Complementary End-to-End 3D human pose generation and estimation Network (CEE-Net) from the aspects of improving the quality of training data and the generalization ability of the pose estimator. More specifically, given a 3D pose $P_{3d} \in \mathbb{R}^{J\times 3}$, with J joints, the generator $\mathcal{G}$ outputs a new 2D-3D pose pairs $\{P^G_{2d} \in \mathbb{R}^{J\times 2}, P^G_{3d} \in \mathbb{R}^{J\times 3}\}$ and feeds them into the estimator $\mathcal{E}$ for training. In order to enable the generator to output more diverse relative 3D pose, we propose the adaptive interpolation sub-network, which can enhance the distribution of positive samples in the discriminator. In addition, our MLP-based body-parts grouping estimator can effectively utilize these new poses, promoting the generator to output more rare poses. The framework is shown in Figure 3.

CEE-Generator

Preliminary In our generator, we use the same Backbone as (Gong, Zhang, and Feng 2021) including several linear layers. When a 3D pose $P_{3d} \in \mathbb{R}^{J\times 3}$ is input into the network, it is transformed into bone vectors $B \in \mathbb{R}^{(J-1)\times 3}$ via a hierarchical transformation $B = \mathcal{H}(X)$ (Wandt and Rosenhahn 2019; Wandt, Ackermann, and Rosenhahn 2018), which can be further decomposed into a bone direction vector $\hat{B}$ and a bone length vector $\|B\|$. Then, the backbone will output a series of coefficients to change the Bone Angle (BA) and Bone Length (BL) of the original pose, and adjust the camera viewpoint through Rigid Transformations (RT) such as rotation and translation to generate new 3D pose P^G_{3d}. Then, the 3D pose can be projected to obtain the corresponding 2D pose P^G_{2d} via the camera parameters.

Joint-Angle limit In (Gong, Zhang, and Feng 2021), the network only uses discriminators to judge the rationality of generated poses, which is unavoidable to output some poses that do not conform to the human body's structural constraints. Therefore, we design the joint-angle loss to make the generated pose within a reasonable range. We denote the joint-angles of the generated pose as $P^G_{3d}(\theta_1, \phi_1, ..., \theta_K, \phi_K)$, and θ_k, ϕ_k represents the kth joint-angle of the pose. Here, we utilize the joint-limit ranges to evaluate the reliability of the generated poses, which are recorded by professional gymnasts wearing sensor devices (Akhter and Black 2015). Then we use joint-limit ranges to detect the reasonable situation of the pose joint angle in P^G_{3d} and use it as supervision.

The local coordinates of these joint-angles can be converted into a spherical coordinate, and the degree of each joint-angle can be mapped on the surface of the sphere. When a joint-angle is outside a reasonable range, the joint-angle loss $\mathcal{L}_\mathcal{J}$ is the shortest distance from the point to the reasonable range on the surface. Otherwise, the loss is 0.

Generation Loss The loss function of our generator is the sum of feedback loss $\mathcal{L}_\mathcal{F}$ (Gong, Zhang, and Feng 2021) and the joint-angle loss:

$$\mathcal{L}_\mathcal{G} = \mathcal{L}_\mathcal{F} + \mathcal{L}_\mathcal{J}, \tag{1}$$

The feedback loss $\mathcal{L}_\mathcal{F}$ can control the complexity of the generated pose according to the error of the estimator.

CEE-Discriminator

Pose augmentation methods usually generate more poses based on the Human3.6M (Ionescu et al. 2013), which is the largest 3D human pose dataset. However, the distribution of poses contained in the dataset is still limited. When it is used as the positive sample in the discriminator, the generator cannot output some reasonable poses that are common but not close to the Human3.6M dataset. Therefore, we aim to increase each joint-angle's distribution range in positive samples of the discriminator.

The simple idea is to take two poses from different actions as the initial and final pose in the source dataset, respectively. Then, each joint-angle is randomly interpolated between the initial pose and the final pose. The pose distribution in the Human3.6M dataset can be expanded through this operation of interpolating bones for different actions. For example, when interpolating between a pose in a standing dataset and a pose in a sitting dataset, a squat pose, which is different from the two actions, will be generated. However, this random approach has no directionality. Therefore, we further propose an adaptive interpolation sub-net to generate a larger distribution of each joint-angle.

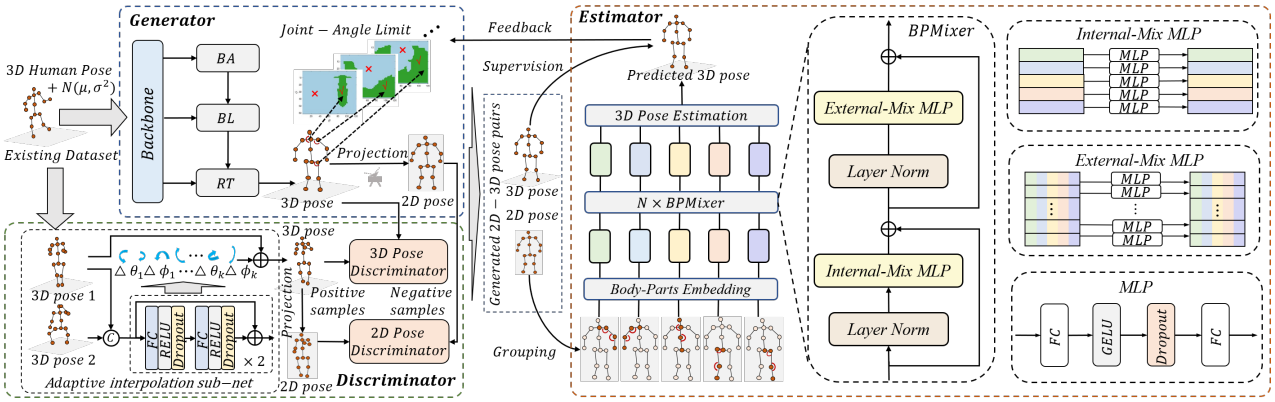

Figure 3: Our end-to-end CEE-Net consists of the generator, discriminator, and estimator in three parts. The generator outputs the new 2D-3D pose pairs by changing the BA, BL, and RT coefficients. At the same time, we add a joint-angle loss to constrain the pose rationality. To generate more diverse poses, we design a adaptive interpolation sub-net to generate additional positive samples for the discriminator. Finally, we propose a new MLP-based body-parts grouping estimator, which has strong robustness to complex poses composed of different body-parts. The generated 2D poses are fed into the estimator, and the output feedback loss can also be used to guide the generator.

The sub-network consists of multiple fully connected layers and residual structures as shown in Figure 3. The input is two-parent skeletons, and the output is the rotation vector of each joint-angle $\Delta\theta_k$ and $\Delta\phi_k$. We use the distribution of the output bone vector for each joint-angle as the supervision, and the loss is inversely proportional to the diversity of the distribution so that the new joint-angles output by the sub-network can contain a larger distribution range and provide more selections for the generator. The visualization of generated positive samples are shown in Figure 5.

Discrimination Loss. For the discrimination loss $\mathcal{L}_{\mathcal{D}}$, We divide the human pose into five body-parts of torso and limbs, and adopt the LS-GAN loss (Mao et al. 2017) for both 3D and 2D spaces:

$$\begin{aligned}\mathcal{L}_{\mathcal{D}} = \sum \{&\mathbb{E}\left[(D_{3d}(B^I_{3d}) - 1)^2\right] + \mathbb{E}\left[(D_{3d}(B^G_{3d}))^2\right] \\ &+\mathbb{E}\left[(D_{2d}(B^I_{2d}) - 1)^2\right] + \mathbb{E}\left[(D_{2d}(B^G_{2d}))^2\right]\},\end{aligned} \tag{2}$$

where $\{B^I_{2d}, B^I_{3d}\}$ denotes the interpolation body-parts pairs (positive smaples), and $\{B^G_{2d}, B^G_{3d}\}$ denote the augmented body-parts pairs (negative smaples), respectively.

CEE-Estimator

In the end-to-end training, when the pose estimation loss is large, a feedback signal will be formed to the generator to attenuate the variation of each joint-angle, which conflicts with improving the diversity of the training set. To solve this problem, we first group joints into different body-parts. As described in SRNet (Zeng et al. 2020), an unseen pose may be composed of several existing poses. This grouping strategy is beneficial to weaken the influence of specific joint-angle changes on the global pose estimation and has strong robustness to complex poses composed of different body-parts. However, the SRNet severely compresses the out-of-group features, easily overfitting the local pose. Recently, (Tolstikhin et al. 2021) propose an all-MLP mixer architecture for visual tasks, which can efficiently combine features within and between different image regions. Inspired by this, we propose a new MLP-based body-parts grouping algorithm with mainly three modules: body-parts embedding, body-parts mixing, and pose estimation. The algorithm focuses on extracting within-group features and effectively combines features of other body-parts outside the group.

Body-Parts Embedding The 2D pose $P^G_{2d} \in \mathbb{R}^{J\times 2}$ output by the generator is used as input to train the estimator. We divide the joints of the 2D pose into five body-parts $B_{2d} \in \mathbb{R}^{5\times 8}$ according to the limbs and torso. Each body-part contains four joints and two joint-angles. Then, each body-part is processed by a learnable embedding, which linearly projects each body-part through a single fully-connected layer to the hidden dimension C. The hidden features can be represented as $H \in \mathbb{R}^{5\times C}$.

Body-parts Mixer The Body-Parts Mixer (BPMixer) is the core part of our estimator, which is used to extract within-group and between-group features of different body parts. The estimator mixing stacks N BPMixer blocks, each consisting of the internal-mix MLP and external-mix MLP two types of MLP layers. The internal-mixing aims to learn spatial dependencies between the joints within the limbs or torso by acting on the rows of the pose embedding H. Each row encodes the spatial features of the same body-part through the layer normalization (Ba, Kiros, and Hinton 2016) and internal-mixing operation, which can be represented as follows:

$$\hat{H} = H + W_2(GeLU(W_1 LayerNorm(H))), \tag{3}$$

where $W_2 \in \mathbb{R}^{D_I\times C}$, $W_1 \in \mathbb{R}^{C\times D_I}$, D_I represents the dimension of internal-mix MLP, $\hat{H}$ represents the internal-mixed features, GeLU represents the Gaussian error linear units activation fuction (Hendrycks and Gimpel 2016).

Since all human limbs can move individually, the internal-mixer MLP makes the estimator robust to complex poses composed of different body parts through grouping training. However, there is an inherently ill-posed problem in the pose transition from 2D to 3D space, which is more obvious when fewer joints are used. Therefore, we propose an external-mixer MLP to alleviate this problem. In order to exchange information between different body-parts, the internal-mixed features $\hat{H}$ are transposed and fed to the external-mixer MLP, and output the external-mixed features $\widetilde{H}$:

$$\widetilde{H} = \hat{H} + (W_4(GeLU(W_3 LayerNorm(\hat{H}^\top))))^\top, \quad (4)$$

where $W_4 \in \mathbb{R}^{D_E \times 5}$, $W_3 \in \mathbb{R}^{5 \times D_E}$, D_E represents the dimension of external-mix MLP. By combining the two MLPs, the estimator can balance the features of the body-parts pose and the global human pose.

3D Pose Estimation In the final prediction stage, the hidden features of all body-parts $\widetilde{H} \in \mathbb{R}^{5 \times C}$ are mapped into a vector of length C through global average pooling. Then, a linear layer is used to regress the length to 64, corresponding to the 3D coordinates of 16 joints in the predicted human pose $\widetilde{P^G_{3d}} \in \mathbb{R}^{J \times 3}$.

Pose estimation loss. The loss function of the pose generator is as follows:

$$\mathcal{L}_{\mathcal{P}} = \left\| P^G_{3d} - \widetilde{P^G_{3d}} \right\|_2, \quad (5)$$

where $\widetilde{P^G_{3d}}$ and P^G_{3d} denote the predicted pose and the ground truth, respectively.

Experiments

Datasets

Human3.6M (Ionescu et al. 2013) contains 3.6 million single-person images and corresponding 3D pose labels. There are 11 actors, including 6 men and 5 women, and a total of 17 action scenes, such as discussion, eating, and greetings. **MPI-INF-3DHP (3DHP)** (Mehta et al. 2017) is a large 3D pose dataset, which covers more diverse human motions than Human3.6M. We use it to evaluate the model's generalization performance. **3DPW** (von Marcard et al. 2018) is a more challenging in-the-wild dataset. It contains more complicated activities and scenarios. Same as 3DHP, we also use it to verify the generalization ability of the proposed methods. **U3DPW** (Li et al. 2020) consists of 300 in-the-wild images with complex human poses, including a variety of extreme sports and unusual actions, and we use it for qualitative analysis.

Implementation Details

The backbone network of the discriminator and generator is consistent with (Gong, Zhang, and Feng 2021). In the adaptive interpolation sub-net, the dimension of FC layers is 1024. In our estimator, hidden dimension C=512, the internal-mix MLP dimension D_I=1024, the external-mix MLP dimension D_E=256. The model first trains the adaptive interpolation sub-net for 20 epochs. Then use new positive samples and train the full network for 50 epochs. The batch size is 1024.

Method	MPJPE	PA-MPJPE
Wang et al. (ECCV'2020) (†)	44.5	34.5
Zeng et al. (ICCV'2021) (†)	45.7	-
PoseFormer (ICCV'2021) (†)	44.3	-
MHFormer (CVPR'2022) (†)	**43.0**	-
Martinez et al. (ICCV'2017)	62.9	47.7
SemGCN (CVPR'2019)	57.6	-
LCN (ICCV'2019)	50.0	40.2
VPose (CVPR'2019)	52.7	40.9
SRNet (ECCV'2020)	49.9	39.4
Zou et al. (ICCV'2021)	49.4	39.1
CEE-estimator	**48.9**	**38.6**
Li et al. (CVPR'2020) (‡)	50.9	38.0
PoseAug (CVPR'2021) (‡)	50.2	39.1
DH-AUG (ECCV'2022) (‡)	49.8	-
CEE-Net (‡)	**47.3**	**36.8**

Table 1: Results on Human3.6M under Protocol #1 and Protocol #2. (†): uses temporal information. (‡): uses pose augmentation.

Method	S1	S1+S5
VPose (CVPR'2019)	65.2	57.9
Li et al. (CVPR'2020) (‡)	61.5	54.6
PoseAug (CVPR'2021) (‡)	56.7	51.3
DH-AUG (ECCV'2022) (‡)	52.2	47.0
CEE-Net (‡)	**51.9**	**46.7**

Table 2: Results on Human3.6M by using varying amounts of training data (S1/S1+S5). We report MPJPE for evaluation.

Comparison with State-of-the-art Methods

Human3.6M We compare the CEE-Net with some state-of-the-art algorithms on the Human3.6M dataset under both Protocol #1 and Protocol #2 (Kanazawa et al. 2018). Following (Gong, Zhang, and Feng 2021), we use the same pose setting (16 joints) and use 2D poses from HR-Net (Sun et al. 2019) as input. The evaluation metrics include MPJPE (Mean Joint Error) and PA-MPJPE (MPJPE after Procrustes alignment). The results are reported in Table 1, which shows that the algorithm using pose augmentation is significantly better than the single pose estimator methods, and our CEE-estimator outperforms other latest pose augmentation methods on two Protocols. Simultaneously, one of the significances of the pose generator is that it can make the pose estimation algorithm obtain high accuracy with less training data. Therefore, we only use S1 and S1+S5 in Human3.6M as training sets to test the effects of different pose generators. Similar to (Lin, Liang, and Deng 2022), we use ground truth as input, and the results are shown in Table 2. The CEE-net shows superior performance over all the other methods, which reflects that our method can obtain competitive results

Method	PCK ↑	AUC ↑	MPJPE ↓
LCN (ICCV'2019)	74.0	36.7	-
HMR (CVPR'2018)	77.1	40.7	113.2
PoseNet 3D (3DV'2020)	81.9	43.2	102.4
Li et al. (CVPR'2020)	81.2	46.1	99.7
RepNet (CVPR'2019)	81.8	54.8	92.5
Wang et al. (ECCV'2020)	84.3	-	90.3
PoseAug (CVPR'2021) (‡)	88.6	57.3	73.0
PoseGU (arXiv'2022) (‡)	86.3	57.2	75.0
DH-AUG (ECCV'2022) (‡)	89.5	57.9	71.2
CEE-Net (‡)	**89.9**	**58.2**	**69.7**

Table 3: Cross-dataset evaluation on 3DHP dataset. PCK (percentage of correct keypoint), AUC (area under the curve) and MPJPE are used for evaluation.

Method	PA-MPJPE
SPIN (ICCV'2019)	96.9
Vpose (CVPR'2019)	94.6
PYMAF (ICCV'2021)	92.8
I2L-MeshNet (ECCV'2020)	93.2
VIBE (CVPR'2020)	93.5
PoseAug (CVPR'2021) (‡)	81.6
PoseGU (arXiv'2022) (‡)	92.3
DH-AUG (ECCV'2022) (‡)	79.3
CEE-Net (‡)	**76.8**

Table 4: Cross-dataset evaluation on 3DPW dataset. PA-MPJPE is used for evaluation.

even using a small amount of training set.

Cross-datasets We use the model trained on the Human3.6M dataset to test the cross-datasets 3DHP and 3DPW to analyze the generalization ability of our end-to-end CEE algorithm. It can be seen from Tables 3 and 4 that all metrics of CEE-Net on the two cross-datasets outperform the previous methods, which indicates that our framework has the best generalization ability. Specifically, our CEE-Net improves MPJPE by 20.6mm on 3DHP and 16.7mm P-MPJPE on 3DPW than the best cross-data pose estimation algorithm without data augmentation, which is a huge improvement. In addition, our complementary end-to-end network improves the augmentation-based state-of-the-art (Lin, Liang, and Deng 2022) from 71.2mm to 69.7mm on the 3DHP dataset and from 79.3mm to 76.8mm on the 3DPW dataset.

Qualitative Results

The end-to-end combination can make the trained estimator suitable for a variety of challenging poses. In order to verify the generalization ability of the CEE-Net on complex poses, we select the extreme pose dataset U3DPW to conduct qualitative analysis with the SOTA pose augmentation method (PoseAug (Gong, Zhang, and Feng 2021)). The experiment results are shown in Figure 4. It can be seen that our network is more effective in dealing with extreme poses.

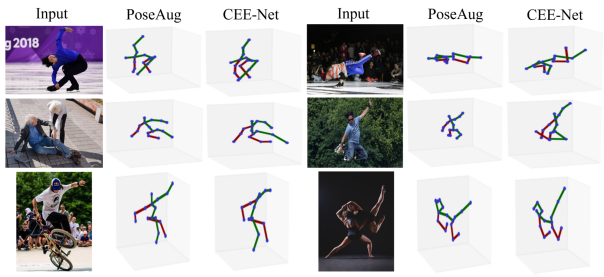

Figure 4: Qualitative comparison results on the complex pose dataset U3DPW with the SOTA pose augmentation method (Gong, Zhang, and Feng 2021).

Ablation Analysis

CEE-Generator in Different Pose Estimator In order to illustrate that CEE-generator can be effectively combined with different 2D pose estimators DET (Girshick et al. 2018), CPN (Chen et al. 2018), HR (Sun et al. 2019) and various 2D-3D regression algorithms SemGCN (Zhao et al. 2019), MLP (Martinez et al. 2017), ST-GCN (Cai et al. 2019), VPose (Pavllo et al. 2019), we conduct comparison experiments on 3DHP, and the detailed results are listed in Table 5. It can be seen that the error is reduced after CEE-generator is combined with any pose estimator than before.

Ablation on BA, BL, and RT As mentioned in the previous sections, although previous data augmentation works have increased the diversity of the training set, the changes in camera viewpoint and position play a dominant role in improving the accuracy of the estimator, while the generated 3D poses are limited. On the contrary, our CEE framework can improve the lack of diversity of 3D pose joint-angles. For effective comparison, we apply our adaptive interpolation sub-net to the original PoseAug and compare the effects of BA, BL, and RT transformations in the two cases. It can be seen from Table 6 that when only BA transformation is included in the pose augmentation, the PoseAug has little performance improvement on the cross-dataset 3DHP, while our performance can significantly improve when only BA changes (5.1mm increase than baseline and 2.9mm than PoseAug). When the three variations of BA, BL, and RT are used in combination, the error of the CEE reaches a 15.2mm decrease from baseline and a 1.6mm from the PoseAug.

Analysis on Complementary In this section, we separately test the effect of the CEE-estimator and CEE-generator in our end-to-end network. Experiments include testing the CEE-estimator alone, using CEE-estimator in combination with other generators, using CEE-generator in combination with other estimators, and our completed CEE-Net. Table 7 summarizes its results on the cross-datasets 3DHP and 3DPW. It is observed that the best performance can be achieved only when our estimator and generator are combined.

Analysis on Joint-Angle Limit As can be seen in Table 7, when the joint-angle limit loss in the generator is removed, the pose estimation error can be decreased from

Method	DET	CPN	HR	GT
SemGCN (CVPR'2019)	101.9	98.7	95.6	97.4
SemGCN + PoseAug	89.9	89.3	89.1	86.1
SemGCN + CEE-generator	**83.6**	**82.8**	**82.4**	**81.3**
MLP (ICCV'2017)	91.1	88.8	86.4	85.3
MLP + PoseAug	78.7	78.7	76.4	76.2
MLP + CEE-generator	**77.6**	**74.9**	**72.7**	**71.8**
ST-GCN (ICCV'2019)	95.5	91.3	87.9	87.8
ST-GCN + PoseAug	83.5	77.7	76.6	74.9
ST-GCN + CEE-generator	**79.5**	**74.8**	**74.6**	**73.4**
VPose (CVPR'2019)	92.6	89.8	85.6	86.6
VPose + PoseAug	78.3	78.4	73.2	73.0
VPose + CEE-generator	**75.6**	**75.2**	**71.2**	**71.4**

Table 5: Performance comparison in MPJPE for various pose estimators trained w/o and with pose generators on 3DHP dataset.

Method	BA	RT	BL	PoseAug	CEE-generator
Baseline				86.6	86.6
Variant A	✓			84.4 (-2.2)	81.5 (-5.1)
Variant B	✓	✓		74.9 (-11.7)	72.2 (-14.4)
Variant C	✓	✓	✓	73.0 (-13.6)	71.4 (-15.2)

Table 6: Ablation study on components of the augmentor for the different discriminator. We report MPJPE for evaluation.

Method	3DHP	3DPW
CEE-Estimator	83.7	91.9
CEE-Generator+Vpose	71.4	78.6
CEE-Estimator+PoseAug	71.8	79.2
Full CEE-Net	**69.7**	**76.8**
w/o joint-angle limit loss	70.4	78.1

Table 7: Ablation study on components of the CEE-Net. PA-MPJPE is used for 3DHP evaluation, and PA-MPJPE is used for 3DPW evaluation.

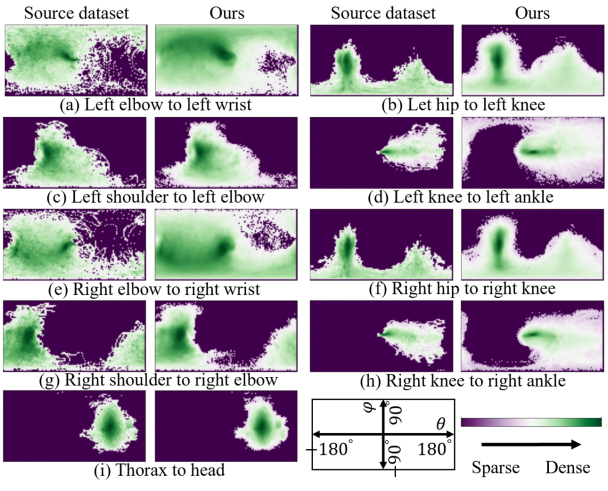

Figure 5: Comparison of each joint-angle's distribution after and before our adaptive interpolation sub-net.

Method	3DHP	3DPW
Random rotation	70.5	77.3
30% rotation	71.6	78.7
50% rotation	71.3	78.2
80% rotation	71.4	78.8
Adaptive Interpolation	**69.7**	**76.8**

Table 8: Ablation study on more interpolation setting in discriminator. PA-MPJPE is used for 3DHP evaluation, and PA-MPJPE is used for 3DPW evaluation.

69.7/76.8mm to 70.4/78.1mm (on 3DHP and 3DPW).

Distribution Visualization In order to more intuitively illustrate the effect of our adaptive interpolation sub-network, we visually compare the distribution of each joint-angle between the original dataset and the generated dataset. As can be seen in Figure 5, we have greatly expanded the distribution range of each joint-angle, which can provide more positive samples for the discriminator and is beneficial for the generator to output more diverse poses.

More Interpolation Settings In the adaptive interpolation sub-net, we try to compare different interpolation methods, including selecting 30%, midpoint, 80%, and a random angle as the transfer joint-angles from the beginning point to the endpoint. The influence of different discriminators on CEE is shown in Table 8. The training-based method has the best effect, while other transfer methods also improve the accuracy to a certain extent, which further illustrates the importance of the positive sample distribution range of the discriminator for the network.

Limitations

By observing all experiment results, it can be found that our CEE-Net has excellent generalization ability in 2D-to-3D pose estimation tasks. However, the 2D pose needs to be extracted from the image in practical applications first. The existing 2D pose estimation algorithms have large errors when dealing with severe occlusions and distorted frames, which will further affect the 3D pose regression.

Conclusion

The paper proposes a novel complementary 3D human pose generation and estimation network to overcome rare and complex pose regression. The complementarity is reflected in that the adaptive interpolation discriminator in our generator can greatly improve the distribution of each joint-angle, providing the estimator with diverse training data, and the body-parts grouping strategy in our estimator can effectively utilize these new poses, promoting the generator to output more rare poses. Compared with other single strong generalization estimators or pose augmentation methods, our architecture can achieve much better performance in multiple cross-datasets and remains stable in the complex pose datasets.

Acknowledgments

This work was supported in part by the Science and Technology Development Fund, Macau SAR, under Grant 0034/2019/AMJ, Grant 0087/2020/A2 and Grant 0049/2021/A.

References

Akhter, I.; and Black, M. J. 2015. Pose-conditioned joint angle limits for 3D human pose reconstruction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1446–1455.

Ba, J. L.; Kiros, J. R.; and Hinton, G. E. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Bouazizi, A.; Holzbock, A.; Kressel, U.; Dietmayer, K.; and Belagiannis, V. 2022. MotionMixer: MLP-based 3D Human Body Pose Forecasting. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, 791–798. International Joint Conferences on Artificial Intelligence Organization.

Cai, Y.; Ge, L.; Liu, J.; Cai, J.; Cham, T.-J.; Yuan, J.; and Thalmann, N. M. 2019. Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2272–2281.

Chen, H.; Tang, H.; Yu, Z.; Sebe, N.; and Zhao, G. 2022. Geometry-contrastive transformer for generalized 3d pose transfer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 258–266.

Chen, T.; Fang, C.; Shen, X.; Zhu, Y.; Chen, Z.; and Luo, J. 2021. Anatomy-aware 3d human pose estimation with bone-based pose decomposition. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(1): 198–209.

Chen, Y.; Wang, Z.; Peng, Y.; Zhang, Z.; Yu, G.; and Sun, J. 2018. Cascaded pyramid network for multi-person pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7103–7112.

Chen, Z.; Huang, Y.; Yu, H.; Xue, B.; Han, K.; Guo, Y.; and Wang, L. 2020. Towards part-aware monocular 3d human pose estimation: An architecture search approach. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, 715–732. Springer.

Duan, H.; Zhao, Y.; Chen, K.; Lin, D.; and Dai, B. 2022. Revisiting skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2969–2978.

Gholami, M.; Wandt, B.; Rhodin, H.; Ward, R.; and Wang, Z. J. 2022. AdaptPose: Cross-Dataset Adaptation for 3D Human Pose Estimation by Learnable Motion Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13075–13085.

Girshick, R.; Radosavovic, I.; Gkioxari, G.; Dollár, P.; and He, K. 2018. Detectron. https://github.com/facebookresearch/detectron. Accessed: 2022-06-23.

Gong, K.; Li, B.; Zhang, J.; Wang, T.; Huang, J.; Mi, M. B.; Feng, J.; and Wang, X. 2022. PoseTriplet: Co-evolving 3D Human Pose Estimation, Imitation, and Hallucination under Self-supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11017–11027.

Gong, K.; Zhang, J.; and Feng, J. 2021. PoseAug: A Differentiable Pose Augmentation Framework for 3D Human Pose Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8575–8584.

Guan, S.; Lu, H.; Zhu, L.; and Fang, G. 2022. PoseGU: 3D Human Pose Estimation with Novel Human Pose Generator and Unbiased Learning. *arXiv preprint arXiv:2207.03618*.

Guzov, V.; Mir, A.; Sattler, T.; and Pons-Moll, G. 2021. Human poseitioning system (hps): 3d human pose estimation and self-localization in large scenes from body-mounted sensors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4318–4329.

Hendrycks, D.; and Gimpel, K. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.

Ionescu, C.; Papava, D.; Olaru, V.; and Sminchisescu, C. 2013. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE TPAMI*.

Kanazawa, A.; Black, M. J.; Jacobs, D. W.; and Malik, J. 2018. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7122–7131.

Kocabas, M.; Athanasiou, N.; and Black, M. J. 2020. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5253–5263.

Kolotouros, N.; Pavlakos, G.; Black, M. J.; and Daniilidis, K. 2019. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2252–2261.

Li, S.; Ke, L.; Pratama, K.; Tai, Y.-W.; Tang, C.-K.; and Cheng, K.-T. 2020. Cascaded deep monocular 3D human pose estimation with evolutionary training data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6173–6183.

Li, W.; Liu, H.; Guo, T.; Tang, H.; and Ding, R. 2022a. GraphMLP: A Graph MLP-Like Architecture for 3D Human Pose Estimation. *arXiv preprint arXiv:2206.06420*.

Li, W.; Liu, H.; Tang, H.; Wang, P.; and Van Gool, L. 2022b. Mhformer: Multi-hypothesis transformer for 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13147–13156.

Lin, Z.; Liang, J.; and Deng, W. 2022. DH-AUG: DH Forward Kinematics Model Driven Augmentation for 3D Human Pose Estimation. In *ECCV*.

Liu, K.; Ding, R.; Zou, Z.; Wang, L.; and Tang, W. 2020. A comprehensive study of weight sharing in graph networks for 3d human pose estimation. In *European Conference on Computer Vision*, 318–334. Springer.

Mao, X.; Li, Q.; Xie, H.; Lau, R. Y.; Wang, Z.; and Paul Smolley, S. 2017. Least squares generative adversarial

networks. In *Proceedings of the IEEE international conference on computer vision*, 2794–2802.

Martinez, J.; Hossain, R.; Romero, J.; and Little, J. J. 2017. A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, 2640–2649.

Mehta, D.; Sridhar, S.; Sotnychenko, O.; Rhodin, H.; Shafiei, M.; Seidel, H.-P.; Xu, W.; Casas, D.; and Theobalt, C. 2017. Vnect: Real-time 3d human pose estimation with a single rgb camera. *ACM Transactions on Graphics (TOG)*, 36(4): 1–14.

Moon, G.; and Lee, K. M. 2020. I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image. In *European Conference on Computer Vision*, 752–768. Springer.

Pavllo, D.; Feichtenhofer, C.; Grangier, D.; and Auli, M. 2019. 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*, 7753–7762.

Shan, W.; Lu, H.; Wang, S.; Zhang, X.; and Gao, W. 2021. Improving robustness and accuracy via relative information encoding in 3d human pose estimation. In *Proceedings of the 29th ACM International Conference on Multimedia*, 3446–3454.

Sun, K.; Xiao, B.; Liu, D.; and Wang, J. 2019. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5693–5703.

Tolstikhin, I. O.; Houlsby, N.; Kolesnikov, A.; Beyer, L.; Zhai, X.; Unterthiner, T.; Yung, J.; Steiner, A.; Keysers, D.; Uszkoreit, J.; et al. 2021. Mlp-mixer: An all-mlp architecture for vision. *Advances in Neural Information Processing Systems*, 34: 24261–24272.

Tripathi, S.; Ranade, S.; Tyagi, A.; and Agrawal, A. 2020. Posenet3d: Learning temporally consistent 3d human pose via knowledge distillation. In *2020 International Conference on 3D Vision (3DV)*, 311–321. IEEE.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

von Marcard, T.; Henschel, R.; Black, M. J.; Rosenhahn, B.; and Pons-Moll, G. 2018. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 601–617.

Wandt, B.; Ackermann, H.; and Rosenhahn, B. 2018. A kinematic chain space for monocular motion capture. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 0–0.

Wandt, B.; and Rosenhahn, B. 2019. Repnet: Weakly supervised training of an adversarial reprojection network for 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7782–7791.

Wang, J.; Yan, S.; Xiong, Y.; and Lin, D. 2020. Motion guided 3d pose estimation from videos. In *European Conference on Computer Vision*, 764–780. Springer.

Wang, Z.; Shin, D.; and Fowlkes, C. C. 2020. Predicting camera viewpoint improves cross-dataset generalization for 3d human pose estimation. In *European Conference on Computer Vision*, 523–540. Springer.

Wehrbein, T.; Rudolph, M.; Rosenhahn, B.; and Wandt, B. 2021. Probabilistic monocular 3d human pose estimation with normalizing flows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 11199–11208.

Xu, F.; Xu, F.; Xie, J.; Pun, C.-M.; Lu, H.; and Gao, H. 2021. Action Recognition Framework in Traffic Scene for Autonomous Driving System. *IEEE Transactions on Intelligent Transportation Systems*.

Xu, T.; and Takano, W. 2021. Graph Stacked Hourglass Networks for 3D Human Pose Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16105–16114.

Yan, S.; Xiong, Y.; and Lin, D. 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Thirty-second AAAI conference on artificial intelligence*.

Zeng, A.; Sun, X.; Huang, F.; Liu, M.; Xu, Q.; and Lin, S. 2020. Srnet: Improving generalization in 3d human pose estimation with a split-and-recombine approach. In *European Conference on Computer Vision*, 507–523. Springer.

Zeng, A.; Sun, X.; Yang, L.; Zhao, N.; Liu, M.; and Xu, Q. 2021. Learning skeletal graph neural networks for hard 3d pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11436–11445.

Zhang, H.; Tian, Y.; Zhou, X.; Ouyang, W.; Liu, Y.; Wang, L.; and Sun, Z. 2021. Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11446–11456.

Zhang, J.; Nie, X.; and Feng, J. 2020. Inference stage optimization for cross-scenario 3d human pose estimation. *Advances in Neural Information Processing Systems*, 33: 2408–2419.

Zhao, L.; Peng, X.; Tian, Y.; Kapadia, M.; and Metaxas, D. N. 2019. Semantic graph convolutional networks for 3d human pose regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3425–3435.

Zheng, C.; Zhu, S.; Mendieta, M.; Yang, T.; Chen, C.; and Ding, Z. 2021. 3d human pose estimation with spatial and temporal transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11656–11665.

Zou, Z.; and Tang, W. 2021. Modulated graph convolutional network for 3d human pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11477–11487.