Shift Equivariant Pose Network

Pengxiao Wang[1], Tzu-Heng Lin[1], Chunyu Wang[3], and Yizhou Wang[1,2]
[1]School of Computer Science, Peking University
[2]Center on Frontiers of Computing Studies, Peking University
[3]Independent Researcher
hbxxwpx@gmail.com, lzhbrian@gmail.com, chunyu.wangdlut@gmail.com, yizhou.wang@pku.edu.cn

Abstract

*Human pose estimation has been greatly advanced in recent years. However, even the best-performing models are not shift equivariant. In particular, a small change in input images often results in drastic alterations in output, which are problematic especially in video applications. The prevalence of top-down approaches, which typically rely on a (non-equivariant) object detector in the first stage, exacerbates this issue. In this paper, we first demonstrate that the biased keypoint representation and the non-equivariant network components are the two main obstacles to shift equivariant pose estimation. To address the limitation, we propose an unbiased decoding method, and redesign the necessary network components (*e.g.*, APS-ResBlock, SSP).Extensive experiments show that our method not only produces much more stable results with shifting input, but also achieves better metrics with the ability of tolerating inaccurate detector output from the first stage. To our knowledge, this is the first work to address the problem of shift equivariance in the field of pose estimation. Our method could be easily applied to existing CNN-based pose estimation networks.*

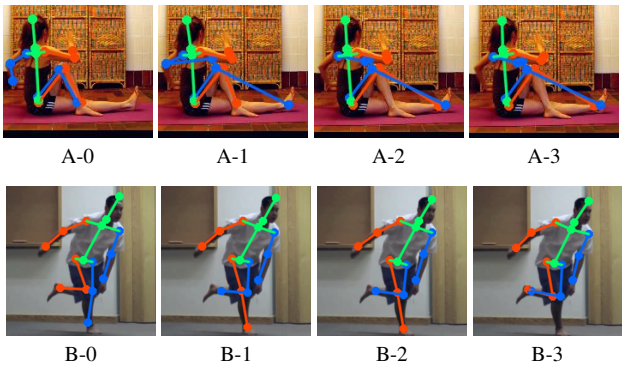

Figure 1. Visualization of the output keypoints of a human pose estimation network [33] when the input is shifted by 0-3 pixels. It is obvious that even a very small input bias can lead to drastically different predictions, e.g., the left wrist in A, and the legs in A and B. Please refer to the videos in the supplementary materials for more visualization.

Table 1. Equivariance Error (**EE**) Comparison on COCO val.

	ViTPose-B	ViTPose-B (w/ hflip TTA)	Ours
$EE_{(1,1)}$	1.267	0.927	0.033

1. Introduction

Human pose estimation has attracted a lot of attention from the computer vision communities, as it is the foundation of many downstream tasks such as action recognition. In recent years, both estimation accuracy and efficiency have been significantly advanced [3, 6, 18, 19, 23, 37] by deep learning. We argue that obtaining stable pose estimation results when input images are slightly perturbed, is equally important as accuracy and speed, as it impacts the user experience to a large extent. Unfortunately, this is largely overlooked in the field. Recently, Chaman et al. [4] pointed out that the existing deep neural networks do not have shift equivariance. In other words, a small disturbance to the input of the network may lead to huge fluctuations in the results. The same problem is also observed in human pose estimation networks. In particular, the popular top-down pose estimators usually rely on an object detector, which is usually not shift equivariant, to detect and crop every person in the first stage. The small localization errors may significantly reduce the accuracy of key point detection. Equivariance is a network characteristic that is **orthogonal** to accuracy. Better equivariance does not guarantee better accuracy (e.g., [4]). While jittering results usually do not harm PCKh@0.5 or AP, it is crucial for applications like **action or medical analysis for athletes or patients**, which requires highly consistent results. Even SOTA model and commonly used Test Time Augmentation (TTA) can not solve the problem of shift equivariance. We provide the Equivariance Error (**EE**) result of ViTPose-B [34], which is

a strong SOTA model. From Table 1, we can see that even one of the best existing models is not shift equivariant at all. Also, TTA (e.g., hflip) cannot solve the problem of shift equivariance.

Fang et al. [10] addressed the issue of pose estimation with inaccurate bounding boxes by introducing a bounding box correction network. However, existing methods fail to perfectly align predicted bounding boxes with the ground truth.An alternative approach is to improve network robustness by achieving shift equivariance, enabling tolerance to input disturbances. Recent studies [4, 38] identify downsampling as the primary cause of broken shift equivariance. Adaptive Polyphase Sampling (APS) [4] was proposed to address this, allowing symmetric convolution networks to maintain shift equivariance by transferring shift information between paired downsampling and upsampling operations.Despite its effectiveness in symmetric networks, APS faces significant limitations. Most pose network architectures are asymmetric for computational efficiency, with input resolutions often exceeding output resolutions, making APS inapplicable. Furthermore, APS fails in scenarios with multiple sampling operations at the same level. Commonly-used blocks, such as ResBlock [12] and DenseBlock [13], lose shift equivariance even with APS. These limitations significantly restrict the practical use of APS in real-world pose estimation tasks.

To address the limitations of shift equivariance in human pose estimation networks, we analyze key impairing factors. Firstly, to mitigate systematic errors from non-model factors, we employ an unbiased Gaussian heatmap coordinate encoding method, necessitating an unbiased decoding approach. We propose **G**aussian **D**istribution prior-based keypoint **P**arameter **E**stimation (**GDPE**), which leverages Gaussian priors to accurately estimate heatmap center coordinates. This method also mitigates quantization errors, ensuring unbiased supervisory signals, consistent with prior work [14]. Secondly, to address errors introduced by model architectures, we propose **APS-ResBlock**, a residual adaptive polynomial sampling block that preserves the consistency of downsampling grids in networks with residual connections. Combined with adaptive polynomial upsampling [4], this ensures shift equivariance during upsampling.

We further resolve boundary errors using circular padding and introduce **S**ubpixel **S**hifting **P**rocessing (**SSP**) to achieve shift equivariance in asymmetric networks. By applying bilinear interpolation on the heatmap in sub-pixel space, SSP enables differentiable shift operations without adding extra parameters.

To summarize, the main contributions of this work are as follows.

- This work introduces shift equivariance into human pose estimation networks for the first time. The proposed method adds no extra learnable parameters and can be generalized to various mainstream network structures.
- We identify biased keypoint representation, including *Coordinate Encoding*, *Coordinate Decoding*, and *Quantization*, as a major factor impairing shift equivariance. To address this, we propose an unbiased coordinate encoding approach and **GDPE**, which accurately estimates keypoint coordinates using Gaussian priors.
- We resolve the shift equivariance issue in asymmetric networks. The proposed **APS-ResBlock** addresses shift non-equivariance caused by multiple sampling operations, while **SSP** enables asymmetric sampling structures to achieve shift equivariance.
- Experiments on **MPII** [1] and **COCO** [21] demonstrate that our method enhances robustness and mitigates accuracy loss caused by bounding box errors, validating its effectiveness.

2. Related Work

Human pose estimation. Human pose estimation task is to locate the key points of the human body in a single image. Recent progress [6, 16, 17, 22, 23, 31, 34] in pose estimation has increasingly improved the accuracy of the system . Fang et al. [10] introduced an extended network to correct the biased bounding box, making the entire network more accurate. Yang et al. [35] used powerful transformer backbone to obtain higher accuracy. Wang et al. [31] fused the local and global features of the input image. Qu et al. [26] used Earth Mover's Distance instead of MSE loss to force the model optimized in a better direction. Li et al. [20] proposed a novel keypoint representation method to replace the classical Gaussian heatmap and achieved better result. UDP [14] and DARK [37] eliminated the keypoint representation error to further improve the accuracy of the syatem.

Shift Equivariance. The success of convolutional neural networks inspired research on embedding equivariances to more complex transformations: rotations, scale, reflections and the action of arbitrary groups [7, 8, 27, 29, 32]. However, the impact of downsampling on the stability of CNNs has only recently been analyzed [2, 4, 9]. Zhang et al. [38] showed that anti-aliasing is able to improve shift invariance in classification. Data augmentation is also proved useful [2]. Chaman et al. [4] proposed APS to enable perfect shift equivariance in symmetric encoder-decoder CNNs.

3. Method

3.1. Preliminaries

Existing human pose estimation networks do not have the property of shift equivariance. Given a system F, shift

equivariance can be modeled as

$$F_\alpha(T_l(I)) = T_{l/\alpha}(F_\alpha(I)), \tag{1}$$

where T represents the translation on the input space and output space, α is the downsampling factor and l is the shift step. As Figure 1 shows, a notable fact is that the keypoints predicted by existing human pose estimation networks [33] are not shifted in the same direction and stride as the input image.

There are two main classes of factors that impair shift equivariance: 1) Non-model factors, which refer to systematic errors (e.g., sometimes even the model's supervisory signal is itself biased). 2) Model factors, which refer to errors introduced by the deep neural network itself (e.g., some basic components of the model may compromise shift equivariance). We propose Unbiased keypoint representation to eliminate the influence of non-model factors. Simultaneously, the structure of shift equivariant pose network is proposed to make the model itself shift equivariant.

3.2. Unbiased keypoint representation

Firstly, we analyze the errors of the existing biased key point representation methods, and point out that there are three systematic errors: Coordinate encoding, Coordinate decoding, and Quantization. Then an unbiased key point representation method is proposed to eliminate these systematic errors introduced by the existing methods.

Coordinate Encoding. The classical coordinate encoding method [33] is to generate a Gaussian heatmap at the label position based on following equations:

$$m' = \begin{cases} \text{Floor}(m) & \text{if } m - \text{Floor}(m) < 0.5 \\ \text{Ceil}(m) & \text{otherwise} \end{cases}, \tag{2}$$

$$n' = \begin{cases} \text{Floor}(n) & \text{if } n - \text{Floor}(n) < 0.5 \\ \text{Ceil}(n) & \text{otherwise} \end{cases}, \tag{3}$$

$$C(x, y, m', n') = \exp(-\frac{(x-m')^2 + (y-n')^2}{2\sigma^2}), \tag{4}$$

where m, n are the horizontal/vertical coordinates of the label, and x, y are the horizontal/vertical coordinates of heatmap. The heatmap representation based on equation 4 is biased. Unbiased coordinate encoding representation shall based on

$$C(x, y, m, n) = \exp(-\frac{(x-m)^2 + (y-n)^2}{2\sigma^2}). \tag{5}$$

We directly generate heatmaps based on Equation 5, without ceil or floor operation. Heatmaps generated by these methods are shown in Figure 2.

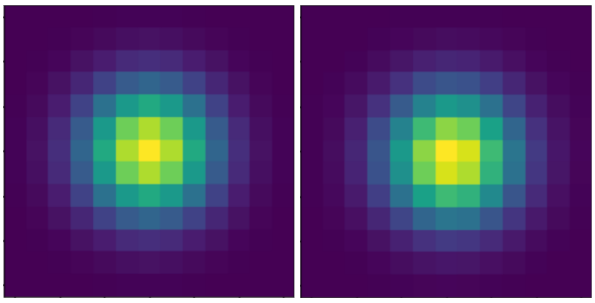

Figure 2. Comparison of heatmaps generated by two coordinate encoding methods when the center coordinate is (6.25, 6.25).The heatmap space is 12×12. Heatmap on the left is generated by biased coordinate encoding methods, the right one is generated by unbiased coordinate encoding methods.

Coordinate Decoding. Based on Equation 4, classical coordinate decoding method [33] is to find the largest corresponding coordinate position of the predicted heatmap. Let c represent the coordinates and the classical coordinate decoding method are as follows:

$$\hat{c} = \arg\max(\hat{C}) + 0.25 * \text{sign}(\hat{C}'(\arg\max(\hat{C}))), \tag{6}$$

$$\text{sign}(x) = \begin{cases} 1 & \text{if } x > 0 \\ -1 & \text{otherwise} \end{cases}. \tag{7}$$

We use $\hat{C}'$ to represent the first derivatives of $\hat{C}$. This coordinate decoding method is also biased. Assuming that the model estimation is completely unbiased, the expected error introduced by this encoding and decoding method: $E(|c_{GT} - \hat{c}|) = 0.125$ [14]. We propose **GDPE** (**G**aussian **D**istribution prior based keypoint **P**arameter **E**stimation) to address the problem above. Assuming that the model can predict perfectly, i.e., $\hat{C} = C$, we expect to decode the label unbiasedly from the unbiased keypoint encoding. This process is the inverse of the encoding process. Given $\{(C_0, x_0, y_0),\ (C_1, x_1, y_1),\ ..., (C_p, x_p, y_p)\}$ and gaussian prior, we need to solve m and n in Equation 5. Taking the logarithm of both sides of Equation 5 at the same time, we get

$$-2\sigma^2 \ln(\hat{C}(x, y, m, n)) = (x-m)^2 + (y-n)^2. \tag{8}$$

Let the objective optimization function be

$$\begin{aligned} F(m, n) = \sum_{i=0}^{p} [&-2\sigma^2 \ln(\hat{C}(x_i, y_i, m, n)) \\ &- (x_i - m)^2 - (y_i - n)^2]^2, \end{aligned} \tag{9}$$

and we use Newton-Raphson method [24] to solve this function. The deductive process is in the appendix. Given an initial value $(m_0, n_0) = \arg\max(\hat{C})$, it is considered

that we can get the exact solution of the function. After multiple iterations, until

$$|\hat{m} - m_0| + |\hat{n} - n_0| < \epsilon, \tag{10}$$

we get the most accurate approximate solution $(\hat{m}, \hat{n})$ to Equation 9.

Quantization. Except for the above two parts, we also need to eliminate the quantization error caused by the classical point-aligned quantization method. Huang et al. [14] had proposed a shift equivariant quantization method. When mapping the annotation information from the image space to the heatmap space, we follow

$$(x', y') = (\frac{x}{\frac{H-1}{h-1}}, \frac{y}{\frac{W-1}{w-1}}), \tag{11}$$

where the height and width of the input image space are (H, W), and that of the output heatmap space are (h, w).

3.3. Shift equivariant pose network

In this section, we introduce the network architecture design. We first analyze the basic components in existing convolutional neural networks: Convolution, Downsampling, Upsampling and Padding operations. We aim to find out the essential factors that destroy the model shift equivariance. Then we introduce basic components of shift equivariant pose network, including APS-U, APS-D, APS-ResBlock, SSP, to address the problem found above. The overall structure of shift equivariant pose network is shown in Figure 3 (a).

3.3.1 Basic Operations

We analyze the basic components and operations in existing convolutional neural networks. Some of them are detrimental to shift equivariance.

Convolution. Convolution can be modeled as:

$$\text{Conv}(I_{i,j}) = \sum_{p,q} I_{i-p,j-q} K_{p,q}. \tag{12}$$

We could prove that the above operation is shift equivariant. Firstly, we formulate shifting operation as:

$$T_{\Delta x, \Delta y}(I_{i,j}) = I_{i+\Delta x, j+\Delta y}. \tag{13}$$

Then we have:

$$\begin{aligned} \text{Conv}(T_{\Delta x,\Delta y}(I_{i,j})) &= \text{Conv}(I_{i+\Delta x,j+\Delta y}) \\ &= \sum_{p,q} I_{i+\Delta x-p,j+\Delta y-q} K_{p,q} \\ &= T_{\Delta x,\Delta y}(\sum_{p,q} I_{i-p,j-q} K_{p,q}) \\ &= T_{\Delta x,\Delta y}(\text{Conv}(I_{i,j})). \end{aligned} \tag{14}$$

Therefore, it is proved that the convolution operation itself is shift-equivariant.

Downsampling. Chaman et al. [4] showed that sampling is one of the main factors that destroy shift equivariance. For the purpose of simplicity, we will consider sampling of 1-D signals $x(n)$. Let U_2 and D_2 denote upsampling and downsampling with stride 2, respectively. D_2 signal is formulated as :

$$D_2(x) = x(2n). \tag{15}$$

In the following, we will prove that linear downsampling does not have shift equivariance. Let $T_\Delta = x(n - \Delta)$ represent a Δ-pixel translation in x. For an odd shift $\Delta = (2m + 1)$ with $m \in \mathbb{Z}$, D_2 satisfies:

$$\begin{aligned} D_2(T_{2m+1}(x)) &= D_2(T_{\lfloor 2m+1 \rfloor}(x)) \\ &= T_{\lfloor \frac{2m+1}{2} \rfloor} D_2(x) \\ &= T_m D_2(x) \\ &= D_2(T_{2m} x) \\ &\neq D_2(T_{2m+1}(x)). \end{aligned} \tag{16}$$

Equation 16 shows that bias can be introduced in linear downsampling when pixels are shifted by an odd number.

Upsampling. Similarly, U_2 signal is formulated as:

$$U_2(x) = \begin{cases} x(n/2) & \text{, when } n \text{ is even} \\ 0 & \text{, otherwise} \end{cases}. \tag{17}$$

The upsampling operation's property of shift equivariance can be easily proved by follows:

$$U_2(T_m(x)) = T_{2m} U_2(x). \tag{18}$$

Padding. Zhang et al. [38] showed that zero padding will cause edge artifacts. During the process of shifting, information is lost on one side and has to be filled in on the other. Circular padding has been proved to solve the above problem [4]. So we use circular padding instead of zero padding in the shift equivariant pose network.

3.3.2 APS-ResBlock

APS-D and APS-U [4] [1] are proposed to make symmetric downsampling and upsampling operations shift equivariant. APS-D chooses the polyphase component of $x(n)$ with the highest l_p norm as its downsampled output $D_2^A(x)$. D_2^A can be formulated as:

$$D_2^A(x) = \begin{cases} x(2n), i_x = 0, & ||x(2n)||_p > ||x(2n+1)||_p \\ x(2n+1), i_x = 1, & ||x(2n)||_p < ||x(2n+1)||_p \end{cases}. \tag{19}$$

[1] A brief introduction of APS is attached in the supplementary materials.

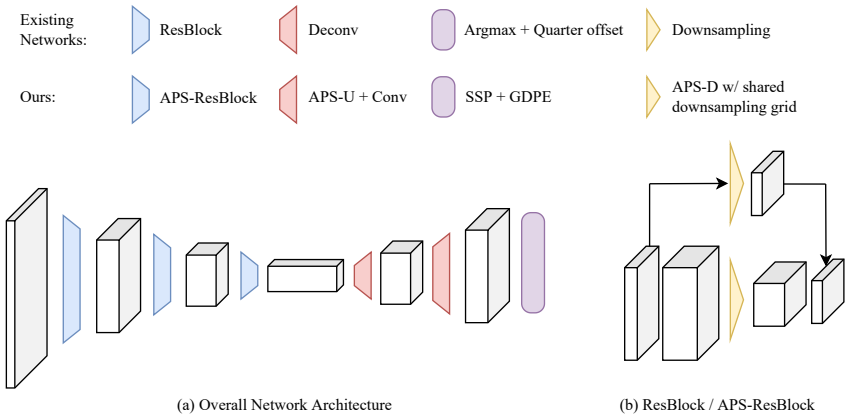

Figure 3. Shift equivariant pose network. APS-Resblock addresses the non-equivariant problem caused by multiple downsampling. Besides, APS-Resblock retains the outstanding feature extraction ability of residual architecture. Existing methods cannot make an asymmetric network shift equivariant. We propose SSP and introduce APS-U to make the pose network shift equivariant for the first time.

We denote i_x to be the index of polyphase component with the highest norm. APS-D is not shift equivariant in the usual sense. However, APS-D can guarantee a consistent output when its input is shifted by odd pixels with i_x to indicate the sampling grid. Then U_2^A takes the sampling grid of the chosen polyphase component i_x as one of the input. U_2^A can be formulated as:

$$U_2^A(x, i_x) = T_{i_x}(U_2(x)). \tag{20}$$

Chaman et al. [4] have proved that

$$U_2^A \circ D_2^A(T_m(x)) = T_m(U_2^A \circ D_2^A(x)), \tag{21}$$

which means a pair of $U_2^A \circ D_2^A$ is shift equivariant.

However, APS method can not be directly applied to human pose estimation networks. Because most of the effective feature extraction blocks, e.g., ResBlock [12], have multiple sampling operations at the same sampling level. Sampling grid can be corrupted when features are merged later. Inside classical residual blocks, each downsampling step actually performs two downsampling operations, which can be described as

$$x_{i+1} = D_2(x_i) + D_2(\text{Conv}(x_i)). \tag{22}$$

Apply APS-D directly to ResBlock, we have

$$x_{i+1} = D_2^A(x_i) + D_2^A(\text{Conv}(x_i)), \tag{23}$$

which means that the sampling grid of $D_2^A(x)$ may not equal $D_2^A(\text{Conv}(x))$. Not surprisingly, these two sampling grid can be corrupted, thus compromising the shift equivariance of the entire system. Inspired by APS, we propose APS-ResBlock to force the entire block to keep the same downsampling grid (i.e., the downsampling grid of $D_2^A(x)$ and $D_2^A(\text{Conv}(x))$ are the same). The structure of APS-ResBlock is shown in Figure 3 (b).

3.3.3 SSP

Another reason APS method can not be directly used in human pose estimation networks is that most pose network structures are not symmetric for the purpose of computational efficiency. As shown in Figure 3, the size of the network's input resolution and output resolution are often different. For the purpose of simplicity, we study a simple asymmetric network. The input image space is (H, W), and the output heatmap space is $(H/4, W/4)$. APS method only works in the case where the output heatmap space is also (H, W). In order to make the asymmetric network shift equivariant, we propose **SSP** (**S**ubpixel **S**hifting **P**rocessing). The structure of **SSP** is shown in Figure 4. If the input is shifted by 1 pixel, network's output can be shifted by 1/4 pixel in the same direction and stride by using SSP. Then this network can be considered shift-equivalent.

SSP operation can be formulated as

$$\text{SSP}(x, i_x) = D_2(T_{i_x}(I(x))), \tag{24}$$

where I denotes bilinear interpolation. Given a downsampling gird with shifting direction and distance information, SSP selects the sub-pixel value obtained by bilinear interpolation and translation as the final heat map output. The SSP operation does not introduce additional errors in other directions or other distances, and it is differentiable. Another advantage of SSP is that SSP does not introduce extra learnable parameters. Using SSP, we can train the shift equivariant network end-to-end without any additional computational cost.

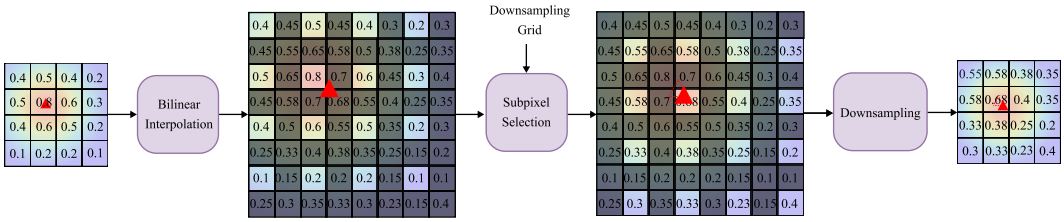

Figure 4. Pipeline of SSP. Through SSP, we achieve shifting heatmap at subpixel level with more precise stride and orientation. For simplicity, we use decimals on the grid to represent pixel values that obey Gaussian distribution. Red triangle denotes the theoretical center of the predicted heatmap. We first perform bilinear interpolation on the heatmap. Then we select and sample the subpixel group based on the downsampling grid from the unpaired D_2^A. Via SSP with $\alpha\times$ interpolation, we can shift the theoretical center of the predicted heatmap at $\frac{1}{\alpha}$ pixel resolution, $\alpha \in \mathbb{Z}$.

4. Experiments

4.1. Experiment Setup

Dataset. We use the MPII dataset [1] and COCO dataset [21] as the experimental dataset. The MPII dataset is collected from YouTube videos with a wide range of human activities and events. It has 25K scene images and 40K annotated persons (29K for training and 11K for test). Each person has 16 labelled body joints. We adopt the standard train/valid/test data split [30]. The COCO keypoint dataset [21] presents naturally challenging imagery data with various human poses. It contains 200k images and 250k person samples. Each person instance is labelled with 17 joints. In evaluation, we follow the commonly used train2017/val2017 split.

Metrics. We use Equivariance Error (**EE**) to measure the shift-equivalent performance of the pose model.

$$\mathrm{EE}_{(\Delta x,\Delta y)} = \frac{1}{N}\sum_{i}^{N} \|T_{\Delta x,\Delta y}(F(I_i)) - F(T_{\Delta x,\Delta y}(I_i))\|. \quad (25)$$

EE aims to calculate the mean squared error between the model output when the input is shifted by $(\Delta x, \Delta y)$ pixels and the shifted model output when the input is not shifted on the entire dataset (with N samples). The smaller the EE index is, the better the shift equivariance of the system is. As the shift becomes larger, a curve with the horizontal axis as the shifting stride and the vertical axis as the EE value can be drawn, which we call the **EE curve**. The smaller the slope of the curve, the more stable the model is against shift perturbations, which means more shift-equivalent properties can be maintained.

Following previous works, for MPII, we use the standard Percentage of Correct Keypoints **PCKh@**τ measurement [36] that quantifies the fraction of correct predictions within an error threshold τ. Specifically, the quantity τ is normalised against the size of head ($\tau = 0.5$, i.e., PCKh@0.5). We measure each individual joint respectively and took their average as an overall metric. For COCO, the standard average precision (**AP**) is used as our evaluation metric, which is calculated based on Object Keypoint Similarity (OKS).

Implementation Details. We implement all experiments in PyTorch [25]. We crop all the training and test images according to the provided positions and scales, and resized them to 256×256 in pixels. As typical, random scaling (0.75-1.25), rotating (±30 degrees) and horizontal flipping were perform to augment the training data.We adopt Adam [15] optimisation algorithm with the following parameter: $\beta_1 = 0.09, \beta_2 = 0.0999, lr = 1e\text{-}3$. The model is trained for a total of 200 epochs, and the learning rate is reduced by 10 times at the 140th, 180th, and 190th epoch, respectively. We use an NVIDIA Titan RTX GPU for training, the overall training time is 22 hours, and the random seed is set to 42.

4.2. Shift Equivariant Experiment

Quantitative results. We show the $\mathrm{EE}_{(1,1)}$ metrics of our method and other baseline methods in Table 2a and Table 2b. It can be seen that by eliminating aliasing, LPF [38] has obtained some improvements in shift equivariance performance. DARK [37] eliminates the systematic errors outside the model and it also strengthens the shift equivariance of the model to a certain extent. In the model where APS [4] is directly applied, the asymmetric structure inside the model is shift equivariant, which also improves the shift equivariance property of the entire model. Finally, our proposed method consistently outperforms the baselines with near-perfect shift equivariance.

Further, in order to measure the preservation range of the model shift equivariance, we draw the EE curve in Figure 5 when the translation is 0-8 pixels. The EE curve of SimpleBaseline [33] shows an increasing trend, indicating that the more the input is shifted, the less equivariant the model output is. It is obvious that SimpleBaseline [33] is not shift equivariant. When even-numbered pixels are shifted, the slope of EE curve will be negative, making the

curve shape jagged. This is caused by systematic errors outside the model, including Coordinate Encoding Error, Coordinate Decoding Error, and Quantization Error. The comparison between DARK [37] and SimpleBaseline [33] can further support the insight that eliminating the effect of systematic errors is beneficial to shift equivariance property. When the systematic errors outside the model are eliminated, EE curve becomes smooth and the slope remains the same. LPF [38] obtains improvements in shift equivariance by eliminating aliasing. The slope of the curve is also reduced. As for APS [4], we directly implement an asymmetric UNet [28] network structure following the method in [4]. As shown in the Figure 5, the model is not shift equivariant when input is shifted by 1-3 pixels. Because APS-D makes the model more inclined to maintain shift invariance when the input is shifted by 1-3 pixels. Our method first removes the quantization error outside the model by GDPE and the slope of the entire EE curve remains consistent. The introduction of APS-ResBlock solves the multiple downsampling problem of the residual network, and the introduction of SSP solves the asymmetry problem of the entire human pose estimation network, so that the entire system is near-perfect shift equivariant.

Qualitative results. Figure 6 visualizes the output of each baseline when the input is shifted by 0-4 pixels. It can be seen that the method in this paper yields a significant improvement in shift equivariance in comparison with other baselines. While other baselines exhibit large deviations in the prediction of the human body structure in case of only a few shifts of input, our method shows excellent consistency. Please refer to the attached videos in the supplementary materials for further comparison.

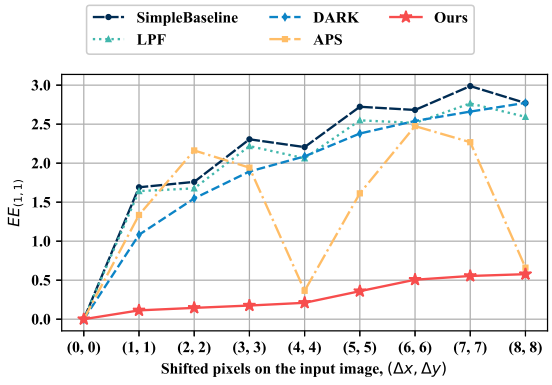

Figure 5. EE curve when input is shifted by 0-8 pixels. The more the input is shifted, the less equivariant the model output is. But our approach has achieved significantly better performance.

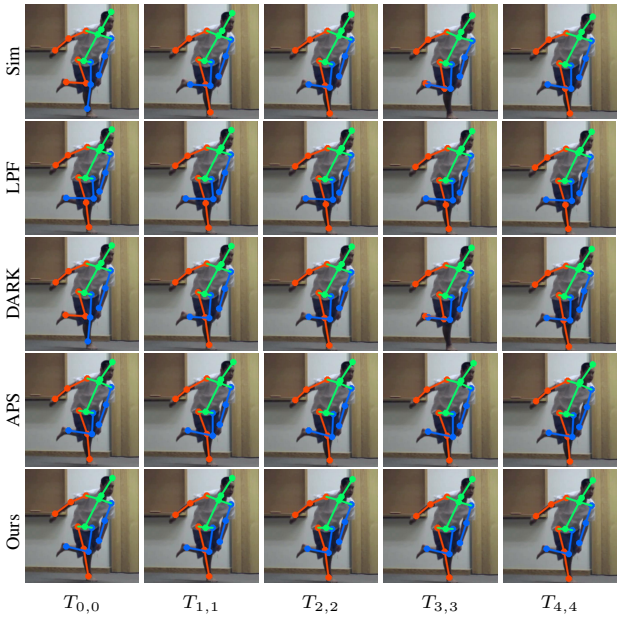

Figure 6. Visualization of the output when the input is shifted by 0-4 pixels for SimpleBaseline [33], LPF [38], DARK [37], APS [4], and Ours. Our method exhibits good equivariance visually. More comparisons are attached in the supplementary materials.

4.3. Model Accuracy Experiment

Table 2a and Table 2b also show the accuracy of the proposed method and the baseline methods on the MPII dataset and COCO dataset. It can be seen that DARK [37] can lead to a small improvement in the accuracy of the model by eliminating the quantization error outside the model. Directly applying APS [4] is incompatible with the pose network's asymmetric architecture, and even result in a relatively large drop. Using LPF [38] to eliminate aliasing can bring about a small increase in model accuracy. Lastly, our method can make the model shift equivariant and obtain similar or higher accuracy than other methods.

4.4. Ablation study

In Table 3, we explore the impact of the key components used in the shift equivariant pose network on the shift equivariance property and prediction accuracy. The introduction of APS-ResBlock leads to a small drop in accuracy, yet the performance of shift equivariance is improved. After separately adding SSP or GDPE, the network can be more shift equivariant and achieve higher accuracy. Finally, our network with all key components above can obtain near-perfect shift equivariance without compromising the accuracy.

We further make a comparison of different decoding methods. The results are shown in Table 4. Our proposed GDPE is the best performing unbiased decoding method with consistent improvements over DARK and UDP. The

Table 2. Quantitative results on MPII and COCO validation dataset. It can be seen that our method achieves a significant improvement in the shift equivariance of the model. Surprisingly, the accuracy of the model does not decrease but increases.

Methods	$EE_{(1,1)}$	PCKh@0.5
Simple. [12]	1.686	87.741
LPF [38]	1.642	88.049
DARK [37]	1.082	88.030
APS [4]	1.336	86.256
Ours	**0.113**	**88.171**

(a) MPII

Methods	$EE_{(1,1)}$	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
Simple. [12]	1.073	68.0	90.6	77.2	65.6	71.9	71.4
LPF [38]	1.014	68.2	90.6	77.2	66.0	71.9	71.7
DARK [37]	0.639	68.7	90.6	77.3	66.2	72.7	72.0
APS [4]	0.894	66.7	88.7	76.0	64.6	72.3	71.7
Ours	**0.033**	**69.2**	**90.6**	**78.3**	**67.3**	**72.0**	**72.5**

(b) COCO validation.

Table 3. Ablation study on 3 key components used in the shift equivariant pose network on MPII.

APS-ResB	SSP	GDPE	$EE_{(1,1)}$	PCKh@0.5
			1.686	87.741
✓			1.244	87.164
✓	✓		1.140	87.653
✓		✓	1.243	87.689
✓	✓	✓	**0.113**	**88.171**

Table 4. Comparison of different decoding methods. Other components are consistent with our final network design (i.e., all of the following lines are equipped with APS-ResBlock, SSP). Argmax* stands for Argmax + QuarterOffset.

Decoding	PCKh@0.5	$EE_{(1,1)}$	$EE_{(3,3)}$	$EE_{(5,5)}$
Argmax*	87.653	1.140	1.227	1.452
DARK	88.087	0.170	0.261	0.513
UDP	88.163	0.148	0.239	0.514
GDPE	**88.171**	**0.113**	**0.176**	**0.359**

Table 5. Inference time per 1k images.

Network	Simple.	LPF	DARK	APS	Ours
second	9.95	9.38	9.80	11.45	19.46

Postprocess	Argmax	DARK	GDPE (5iter)
second	14.95	45.39	82.26

greater the input deviation is, the more obvious the shift equivariance advantage of GDPE is.

Inference speed is also an important consideration in applications. We report the time cost for both network part and postprocessing part in Table 5. For network latency, our model has reached a speed that can be used for real-time inference. For postprocessing, we could choose GDPE if higher consistency are required, and slower speed can be tolerated. We leave accelerating the inference speed as an important future work.

4.5. End-to-end Robustness Experiment

It is hoped that the shift equivariant pose network can eliminate the performance impact of the shifting jitter of the front-connected human detector. Following the real usage scenario, we use Yolo-X [11] as the front-connected human detector, and input the detected human bounding box into the subsequent human pose estimation network. The numerical results of the experiments are shown in Table 6. We can see that the method proposed in this paper outperforms the baseline model. This shows that the shift equivariant pose network can effectively alleviate the detection deviation problem. This is of great significance in practical application.

Table 6. Accuracy (PCKh@0.5) with Yolo-X prepended on MPII.

Method	GT bbox	Yolo-X bbox	gap
DARK [37]	88.030	84.457	3.573
Ours	**88.171**	**85.496**	**2.675**

5. Conclusion

Existing pose estimation models lack shift equivariance, leading to significant output jitter with small input shifts. This issue is further aggravated by non-equivariant detector outputs. In this paper, we introduce shift equivariance into human pose estimation networks for the first time. Extensive experiments demonstrate that our method produces more stable and accurate results under shifting inputs and improves tolerance to inaccurate detector outputs, achieving better overall metrics. However, our method cannot yet be applied to transformer-based models due to the non-equivariant nature of MLPs [5]. Additionally, current approaches do not address rotation and scaling equivariance, which are crucial for practical applications. Future work will focus on integrating rotation and scaling equivariance into pose estimation networks to enhance accuracy and robustness.

6. Acknowledgments

This work was supported in part by the National Science and Technology Major Project (2022ZD0114904). The authors would like to thank them for their valuable contributions.

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, pages 3686–3693, 2014. 2, 6

[2] Aharon Azulay and Yair Weiss. Why do deep convolutional networks generalize so poorly to small image transformations? *arXiv preprint arXiv:1805.12177*, 2018. 2

[3] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017. 1

[4] Anadi Chaman and Ivan Dokmanić. Truly shift-equivariant convolutional neural networks with adaptive polyphase upsampling. *arXiv preprint arXiv:2105.04040*, 2021. 1, 2, 4, 5, 6, 7, 8

[5] Anadi Chaman and Ivan Dokmanic. Truly shift-invariant convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3773–3783, 2021. 8

[6] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7103–7112, 2018. 1, 2

[7] Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International conference on machine learning*, pages 2990–2999. PMLR, 2016. 2

[8] Taco S Cohen, Mario Geiger, Jonas Köhler, and Max Welling. Spherical cnns. *arXiv preprint arXiv:1801.10130*, 2018. 2

[9] Logan Engstrom, Brandon Tran, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. Exploring the landscape of spatial robustness. In *International conference on machine learning*, pages 1802–1811. PMLR, 2019. 2

[10] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. Rmpe: Regional multi-person pose estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 2334–2343, 2017. 2

[11] Z Ge, S Liu, F Wang, Z Li, and J Sun. Yolox: Exceeding yolo series in 2021. arxiv. *arXiv preprint arXiv:2107.08430*, 2021. 8

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2, 5, 8

[13] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 2

[14] Junjie Huang, Zheng Zhu, Feng Guo, and Guan Huang. The devil is in the details: Delving into unbiased data processing for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5700–5709, 2020. 2, 3, 4

[15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[16] Sven Kreiss, Lorenzo Bertoni, and Alexandre Alahi. Pifpaf: Composite fields for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11977–11986, 2019. 2

[17] Jia Li, Wen Su, and Zengfu Wang. Simple pose: Rethinking and improving a bottom-up approach for multi-person pose estimation. In *Proceedings of the AAAI conference on artificial intelligence*, pages 11354–11361, 2020. 2

[18] Ke Li, Shijie Wang, Xiang Zhang, Yifan Xu, Weijian Xu, and Zhuowen Tu. Pose recognition with cascade transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1944–1953, 2021. 1

[19] Wenbo Li, Zhicheng Wang, Binyi Yin, Qixiang Peng, Yuming Du, Tianzi Xiao, Gang Yu, Hongtao Lu, Yichen Wei, and Jian Sun. Rethinking on multi-stage networks for human pose estimation. *arXiv preprint arXiv:1901.00148*, 2019. 1

[20] Yanjie Li, Sen Yang, Shoukui Zhang, Zhicheng Wang, Wankou Yang, Shu-Tao Xia, and Erjin Zhou. Is 2d heatmap representation even necessary for human pose estimation? *arXiv preprint arXiv:2107.03332*, 2021. 2

[21] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 2, 6

[22] Zhengxiong Luo, Zhicheng Wang, Yan Huang, Liang Wang, Tieniu Tan, and Erjin Zhou. Rethinking the heatmap regression for bottom-up human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13264–13273, 2021. 2

[23] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, pages 483–499. Springer, 2016. 1, 2

[24] Isaac Newton. *The Method of Fluxions and Infinite Series: With Its Application to the Geometry of Curve Lines*. Nourse, 1736. 3

[25] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 6

[26] Haoxuan Qu, Li Xu, Yujun Cai, Lin Geng Foo, and Jun Liu. Heatmap distribution matching for human pose estimation. *arXiv preprint arXiv:2210.00740*, 2022. 2

[27] Siamak Ravanbakhsh, Jeff Schneider, and Barnabas Poczos. Equivariance through parameter-sharing. In *International conference on machine learning*, pages 2892–2901. PMLR, 2017. 2

[28] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 7

[29] Ivan Sosnovik, Michał Szmaja, and Arnold Smeulders. Scale-equivariant steerable networks. *arXiv preprint arXiv:1910.11093*, 2019. 2

[30] Jonathan Tompson, Ross Goroshin, Arjun Jain, Yann LeCun, and Christoph Bregler. Efficient object localization using convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 648–656, 2015. 6

[31] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364, 2020. 2

[32] Maurice Weiler, Fred A Hamprecht, and Martin Storath. Learning steerable filters for rotation equivariant cnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 849–858, 2018. 2

[33] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 466–481, 2018. 1, 3, 6, 7

[34] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *arXiv preprint arXiv:2204.12484*, 2022. 1, 2

[35] Sen Yang, Zhibin Quan, Mu Nie, and Wankou Yang. Transpose: Keypoint localization via transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11802–11812, 2021. 2

[36] Yi Yang and Deva Ramanan. Articulated human detection with flexible mixtures of parts. *IEEE transactions on pattern analysis and machine intelligence*, 35(12):2878–2890, 2012. 6

[37] Feng Zhang, Xiatian Zhu, Hanbin Dai, Mao Ye, and Ce Zhu. Distribution-aware coordinate representation for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7093–7102, 2020. 1, 2, 6, 7, 8

[38] Richard Zhang. Making convolutional networks shift-invariant again. In *International conference on machine learning*, pages 7324–7334. PMLR, 2019. 2, 4, 6, 7, 8