Robust Human Pose Estimation for Rotation via Self-Supervised Learning

KIMIN YUN1, JONGYOUL PARK1, AND JUNGCHAN CHO2, (Member, IEEE)

1Artificial Intelligence Research Laboratory, Electronics and Telecommunications Research Institute, Daejeon 34129, South Korea
2Department of Software, Gachon University, Seongnam 13120, South Korea

Corresponding author: Jungchan Cho (thinkai@gachon.ac.kr)

This work was partly supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. B0101-15-0266, Development of High Performance Visual BigData Discovery Platform for Large-Scale Realtime Data Analysis and No. 2020-0-00004, Development of Previsional Intelligence based on Long-term Visual Memory Network).

ABSTRACT The detection of abnormal postures, such as that of a reclining person, is a crucial part of visual surveillance. Further, even regular poses can appear rotated because of incongruity between the image and the angle of a pre-installed camera. However, most existing human pose estimation methods focus on small rotational changes, i.e., those less than 50 degrees, and they seldom consider robust human pose estimation for more drastic rotational changes. To the best of our knowledge, there have been no reports on the robustness of human pose estimation for rotational changes through large angles. In this study, we propose a robust human pose estimation method by creating a path for learning new rotational changes based on a self-supervised method and by combining the results with those obtained from a path based on a supervised method. Furthermore, a combination module composed of a convolutional layer is trained complementarily by both paths of the network to produce robust results for various rotational changes. We demonstrate the robustness of the proposed method with extensive experiments on images generated by rotating the elements of standard benchmark datasets. We fully analyze the rotational characteristics of the state-of-the-art human pose estimators and the proposed method. On the COCO Keypoint Detection dataset, the proposed method attains more than 15% improvement in the mean of average precision compared to the state-of-the-art method, and the standard deviation of the performance is improved by more than 4.7 times.

INDEX TERMS Deep learning, human pose estimation, rotation, self-supervised learning.

I. INTRODUCTION

The problem of estimating a human pose based on a single image can be reduced to accurately locating a set of semantic keypoints, e.g., the head, shoulder, elbow, wrist, knee, and ankle of a human body [1]. This is a fundamental task in more complicated problems such as action recognition [2], [3], and it is useful in various applications such as human-computer interaction and animation [4], [5].

Owing to the importance of the problem, human pose estimation has been extensively studied in various fields [1], [2], [6]–[13]. We focus on single person pose estimation because it is the basic step in other applications, such as multi-person pose estimation, video-based human pose estimation, and tracking in two-dimensional or three-dimensional spaces.

The associate editor coordinating the review of this manuscript and approving it for publication was Huanqiang Zeng.

Traditional methods [14], [15] that rely on hand-crafted features are limited in performance owing to the large variability in the appearance of the human body, which can change considerably because of alterations in light conditions and human movement. In addition, complex backgrounds and occlusion problems are also challenges to this task. Recently, deep learning has produced successful results in various tasks related to computer vision, including classification and recognition of images [16], [17]. The field of human pose estimation is no exception [18]. Through the efforts invested by many researchers and the emergence of large datasets such as the COCO dataset [19], the performance of human pose estimation seems to have reached a practically sufficient level [7], [13], [20], [21]. Nevertheless, some problems remain to be solved to ensure wide-ranging practical applications.

We contend that estimating a rotated human pose through a large angle is one of the main unaddressed issues. To the
To demonstrate the robustness of the proposed human pose estimation method, extensive experiments were conducted using input images generated by rotating images in the MPII [24] and COCO [19] datasets. The experimental results verify the superior performance of the proposed method over existing state-of-the-art methods [20], [21]. Further, we also demonstrate qualitatively the improved results of the proposed method on several datasets [25]–[27] that contain images of people in reclining postures because of the angle of the pre-installed camera.

The contributions of this paper are as follows.

- We fully analyze the rotational characteristics of the state-of-the-art human pose estimators [20], [21]. Thereby, we show that the existing methods, including the state-of-the-arts [20], [21], are not robust in the case of large rotational changes. This leads us to the need for research on robust methods for large rotational changes.

- We propose a robust human pose estimation method by creating a new path for learning rotational changes based on a self-supervised method and combining it with the results of a path based on a supervised method. Further, a combination module with multiple paths is complementarily trained via a convolutional layer of the network to produce robust results for various rotational changes.

The remainder of this paper is organized as follows: Section II describes the related works on human pose estimation based on convolutional neural networks (CNNs) and self-supervised methods for various visual tasks. Section III illustrates the proposed combination of supervised and self-supervised methods. The detailed implementations and results of the conducted experiments are presented in Sections IV and V, respectively. Finally, Section VI concludes the paper.

II. RELATED WORK

A. HUMAN POSE ESTIMATION

In 2012, AlexNet [29], which utilized a convolutional neural network (CNN), won the ILSVRC (ImageNet Large Scale Visual Recognition Challenge) with an overwhelming difference in the quality of performance compared to other methods that did not use CNNs. Since then, many studies in the field of computer vision have utilized deep learning technology [17], which has led to significant improvements in human pose estimation accuracy [18].

Chen et al. [12] proposed a method called the cascaded pyramid network (CPN), which integrates all levels of feature representation to alleviate difficulties in estimating “hard” keypoints such as those that are occluded or invisible. Xiao et al. [20] observed that many human pose estimation algorithms are too complex, and they provided a simple baseline method (SimpleBaseline) that is based on a few deconvolutional layers added on a backbone network, ResNet [17]. However, these methods utilize low-resolution representations obtained from feature extraction networks to recover high-resolution representations. This is a bottleneck...
that makes it difficult to obtain rich high-resolution representations for accurate and precise human pose estimation. To overcome this limitation, Sun et al. [21] focused on the process of learning rich high-resolution representations (HRNet) by designing a multi-scale, multi-stage network. HRNet starts with a high-resolution sub-network and gradually adds parallel low-resolution sub-networks. It then repeatedly performs multi-scale fusion to obtain rich high-resolution representations. Further, Sun et al. empirically demonstrated the effectiveness of the multi-stage network based on its superior human pose estimation results compared to that of standard benchmark datasets [19], [24].

Despite the success of such studies, our analysis in Section V demonstrates that even state-of-the-art methods [20], [21] are vulnerable to inaccuracies arising from large rotational changes.

B. SELF-SUPERVISED LEARNING

Recently, studies on self-supervised learning methods have overcome the limitation of intensive manual labeling efforts required in supervised learning. Gidaris et al. [10] attempted to predict image rotations by using a self-supervised method, and then, they demonstrated that the learned features could be utilized in various tasks. Zhang et al. [22] proposed a method that learned image representations using a CNN to recognize geometric transformations applied to an input image. The deep network first randomly generated a small set of discrete geometric transformations and applied each of those geometric transformations to each image on a dataset. Then, the transformed images are transmitted to the CNN trained to recognize the transformations of each image. To estimate the transformation, important feature representations needed to be learned in an unsupervised manner. Feng et al. [23] proposed a method to learn split representations containing parts both related and unrelated to the rotation of a posture in an image. By decoupling the rotation from instance discrimination, their method became the state-of-the-art in terms of standard self-supervised feature learning benchmarks.

The aforementioned methods aim to perform comparably to supervised methods with greater ease and lesser data and effort by learning features in a self-supervised manner. However, the method proposed in this paper is fundamentally different from the aforementioned methods because it aims to overcome the limitation of supervised methods by combining the results obtained from supervised and self-supervised paths, as depicted in Figure 2.

III. APPROACH

As described in Section I, a robust human pose estimator that considers rotational changes cannot be easily designed using supervised learning. To overcome this limitation, the proposed method uses two paths: an original supervised path (the upper path in Figure 2) and a new path (the lower path in Figure 2) for learning, which is complementary to the rotational changes. Both paths are trained by sharing weights through a Siamese network structure as depicted in Figure 2. To simulate the complementary rotational changes in a self-supervised manner, we apply a rotational transformation on the training data using a spatial transformer [30]. Then, we combine the output feature map from the self-supervised path with the corresponding supervised one. A combination module that comprises a convolutional layer is designed to learn how to produce robust results for various rotational changes.

A. REVIEW OF A SPATIAL TRANSFORMER

Jaderberg et al. [30] introduced a spatial transformer based on a CNN. It is a differential module for the spatial transformation of an input feature map during a single forward pass. For simplicity, this section only considers a single channel input and a single transform. However, the method can be generalized to multiple-channel inputs and multiple transformations.
For multi-channel inputs, the same warping is applied to each channel, as depicted in [30]. The spatial transformer mechanism is split into three parts (a localization network, grid generator, and sampler):

1) The localization network \( f_{loc} \) estimates the set of parameters \( \Theta \) of a spatial transformation \( T_{\Theta}(U) \) of an input feature map \( U \in \mathbb{R}^{H \times W \times C} \), which can be represented as \( \Theta = f_{loc}(U) \), where \( H \) and \( W \) are the height and width of an input (source) grid \( G' \), and \( C \) is the number of channels. The size of \( \Theta \) can vary depending on a parameterized transformation type, e.g., for a 2D affine transformation, \( \Theta \) is six-dimensional.

2) The grid generator creates an input sampling grid \( G' \) by warping a regular (target) grid \( G' \) based on the predicted transformation parameters, i.e., \( G' = T_{\Theta}(G') \).

In general, the output elements of a feature map are defined to lie on a regular grid \( G' \) of elements \( G'_i = (x'_i, y'_i) \). For instance, if \( T_{\Theta} \) is a 2D affine transformation matrix of \( \mathcal{A}_\Theta \), then the point-wise transformation is

\[
\begin{bmatrix}
    x'_i \\
    y'_i 
\end{bmatrix} = T_{\Theta}(G'_i) = \mathcal{A}_\Theta \begin{bmatrix}
    x'_i \\
    y'_i \\
    1
\end{bmatrix} = \begin{bmatrix}
    \Theta_{11} & \Theta_{12} & \Theta_{13} \\
    \Theta_{21} & \Theta_{22} & \Theta_{23}
\end{bmatrix} \begin{bmatrix}
    x'_i \\
    y'_i \\
    1
\end{bmatrix}.
\]

Note that \((x'_i, y'_i)\) and \((x'_i, y'_i)\) are the coordinates that define the \( i \)-th element in the input and output feature maps, respectively. The transformation and sampling are equivalent to the standard texture mapping and coordinates used in graphics [30], [31]. The transformation can have any parameterized form if it is differentiable with respect to the parameter [30].

3) The sampler generates a transformed output feature map \( V \in \mathbb{R}^{H' \times W' \times C'} \) sampled from the input feature map \( U \) at the grid points \( G' \), where \( H' \) and \( W' \) are the height and width of the output grid \( G' \), respectively, and \( C' \) is the number of channels, which is the same for the input and output.

In this paper, we assume that transformation parameters are given, and that the transformation is rotational in order to generate training data. Then, we perform an inverse transformation of the output human pose feature map and the estimated heatmap of the self-supervised learning path, as depicted in Figure 2. In other words, unlike [30], the proposed method only uses a grid generator and a sampler alongside the given transformation parameter, \( \Theta \). In this case, the rotation module is differentiable, and therefore, the proposed network becomes an end-to-end trainable deep neural network.

For the convenience of explanation, the input of the spatial transformer function will be referred to as an image (feature map) instead of a grid, and the transformed image (feature map) will be referred to as the output in Sections III-B and III-C.

**B. SELF-SUPERVISED LEARNING WITH A SET OF ROTATED IMAGES**

The aim of this paper is to train a CNN based on robust rotation features using a self-supervised method based on a set of \( N \) input images complementary to each other with respect to the rotations.

**Definition 1 (Complementary Image Set):** Let \( I_0 \) be an original input image with zero rotational angle, and let \( I_i \) be the \( i \)-th rotated input image based on \( I_0 \) with a rotational angle, \( \theta \). Then, the set of complementary images corresponding to ratio through the angle, \( \theta \), is \( C_\theta = \{I_i\} \), where the \( i \)-th image \( I_i \) has the following relationship with \( I_0 \).

\[
I_i = T_{\theta^i}(I_0),
\]

where \( T_{\theta^i}(\cdot) \) denotes the \( i \)-fold application of a rotational transformation through an angle, \( \theta \).

As the rotated image \( I_i \) (if \( i \neq 0 \)) can be easily generated by a spatial transformer in a self-supervised manner, for robust human pose estimation corresponding to rotational changes, we can consider an objective function to reduce the loss on complementary image sets.

Let \( p_0 \) and \( p_i \) be the ground truth human poses corresponding to \( I_0 \) and \( I_i \), respectively. Then, \( p_i = T_{\theta^i}(p_0) \) and a loss function for training a complementary image set can be taken to be

\[
\sum_{i=1}^{\left|C_\theta\right|} \text{loss}(F(I_i), p_i),
\]

where \( F(\cdot) \) is a human pose estimator and \( \text{loss}(\cdot) \) captures general human pose estimation loss, such as the weighted mean squared error discussed in [20], [21].

However, as is clear from Section V, simple data augmentation for rotational changes is not sufficient to ensure a robust human pose estimator for large rotational changes.

**C. COMBINATION MODULE FOR A ROBUST HUMAN POSE ESTIMATION RESULT**

We address this problem by devising a module that combines the results obtained from supervised learning and self-supervised learning paths. In particular, if \( \left|C_\theta\right| \) is two, then a Siamese network can be devised, as is depicted in Figure 2:

- First, the proposed network itself generates an additional rotated input image by rotating the input image through a large angle.

- Then, the original and rotated human poses are estimated simultaneously. Here, the second path that estimates the rotated human pose based on a rotated image can be considered as a type of self-supervised learning path because the proposed network itself produces rotated training images and their corresponding ground truths.

- Finally, we use a combination module based on a convolutional operator that produces the final human pose estimation result.
To achieve this goal, the overall objective function is formulated as
\[
\sum_{i=1}^{C_i} \text{loss}(\mathcal{T}_\theta^{-1}(F(I_i)), p_0) + \text{loss}(H(+\sum_{i=1}^{C_i} \mathcal{T}_\theta^{-1}(\tilde{F}(I_i))), p_0),
\]
where \(\tilde{F}(\cdot)\) is a human pose feature extractor, which produces the output feature map immediately preceding the final layer of human pose estimator, and where the final layer matches the number channels of human pose features to the number of human joints. Further, \(+\) is a channel concatenation operator and \(H(\cdot)\) is a combination module that consists of a convolutional layer. Any conventional loss for human pose estimation can be used for \(\text{loss}(\cdot)\), such as the weighted mean squared error discussed in [20], [21]. As is apparent from the definition of the complementary image set and the form of the objective function, the independent loss term, i.e., \(\sum_{i=1}^{C_i} \text{loss}(\mathcal{T}_\theta^{-1}(F(I_i)), p_0)\), learns the complementary rotations with respect to each other, and the latter combination loss term, i.e., \(\text{loss}(H(+\sum_{i=1}^{C_i} \mathcal{T}_\theta^{-1}(\tilde{F}(I_i))), p_0)\), combines this complementary information to produce a robust result, as demonstrated in Section V.

IV. IMPLEMENTATION

Before discussing the experiments, we describe the parameters and settings for training convolutional neural networks, an evaluation protocol, and the evaluation datasets.

A. TRAINING SETUP

To ensure a fair comparison between the two baseline methods [20], [21] and the proposed method, we used almost the same set of parameters and settings as the one discussed in SimpleBaseline [20] and HRNet [21]. The learning rate was adjusted using the Adam Optimizer [32]. Using the backbone structure of SimpleBaseline [20], the base learning rate was initialized at 1e-3 and dropped to 1e-4 and 1e-5 after 90 and 120 epochs, respectively, and the CNN was trained for 140 epochs. In the case of HRNet backbone [21], the base learning rate was initialized at 1e-3, dropped to 1e-4 and 1e-5 after 170 and 200 epochs, respectively, and the CNN was trained for 210 epochs. Both methods utilized the ImageNet pretrained backbone network: SimpleBaseline used the ResNet series [17] and HRNet [21] used its own HRNet as a backbone network.

The same data augmentation was applied to train baseline methods used in [20], [21]. For the MPII dataset [24], random rotations through \([-30, 30]\) degrees, random scalings in \([0.75, 1.25]\), and horizontal flips were applied.\(^1\) For the COCO dataset [19], random rotations through \([-45, 45]\) degrees, random scalings in \([0.65, 1.35]\), and horizontal flips were applied.

In the proposed method, only one data augmentation was added to emulate input images depicting reclining postures: an image rotated through -90 degrees was augmented with 50% probability for a supervised path image. However, we compared the proposed method with the results obtained from training via expanding the rotation augmentation to \([-90, 90]\) degrees in Section V-B. The cardinality of the compilation image set for the proposed method was two, i.e., \(C_{90} = \{I_0, I_1\}\), and the image corresponding to the self-supervised path was an image rotated through 90 degrees, i.e., \(I_1 = \mathcal{T}_{90}(I_0)\).

B. MPII HUMAN POSE ESTIMATION

The MPII human pose dataset [24] is a dataset of various human daily activities with the coordinates and visibilities of 16 human joints. This dataset comprises over 25K images and 40K people. The dimensions of the input image were cropped to \(256 \times 256\), to provide a fair comparison with the other methods.

For the evaluation metric, we used the PCKh (head-normalized probability of correct keypoint) score, which is the standard metric in MPII human pose estimation [24]. Among them, PCKh@0.5 is used for joint localization accuracy, which matches a joint if the distance between the estimated joint point and the ground-truth is less than 0.5 times the length of the head segment.

C. COCO KEYPOINT DETECTION

The COCO dataset [19], widely used in object detection, records images and labels for keypoint detection. This dataset is labeled with 17 keypoints and visibilities for over 150K people. As the existing methods [20], [21] evaluated the performance on \(256 \times 192\) images by cropping the heights and widths in a 4:3 ratio, we re-trained these baseline networks to utilize \(256 \times 256\) input images to ensure a fair comparison.

The mean average precision (AP) and average recall (AR) were used as the evaluation metrics based on object keypoint similarity (OKS), which is the standard metric in the COCO keypoint detection challenge.\(^2\) OKS is a measure that converts the Euclidean distance \(d^2\) between the ground truth keypoint and the estimated keypoint to a value between 0 and 1, as follows.

\[
\text{OKS} = \frac{\sum_i \exp(-s^2/2s^2k^2)\delta(v_i > 0)}{\sum_i \delta(v_i > 0)},
\]

where \(s^2\) is a scale parameter, and \(k^2\) and \(v_i\) indicate a falloff control parameter and the visibility of the \(i\)-th joint, respectively. AP\(^{20}\) (the average precision at OKS = .50), AP\(^{75}\) (the average precision at OKS = .75), AP (the mean of AP scores at OKS = 0.5, 0.55, 0.6, . . . , 0.95), and AR (the mean of average recall scores at OKS = 0.5, 0.55, 0.6, . . . , 0.95)

\(^1\) The random rotation \(r\) in \([-30, 30]\) means that the rotation is sampled from a Gaussian distribution with the zero mean and 30 degrees standard deviation, i.e., \(r \sim \mathcal{N}(0, 30^2)\). The random scaling \(s\) in \([0.75, 1.25]\) indicates that the scaling factor is sampled from a Gaussian distribution with one mean and 0.25 standard deviation, i.e., \(s \sim \mathcal{N}(1, 0.25^2)\). That is, about 68% of rotation and scaling parameters are sampled from those ranges. In this paper, the square bracket is defined for representing data augmentation parameters.

\(^2\) http://cocodataset.org/#keypoints-eval
FIGURE 3. Evaluation Protocol for an example image in the COCO validation dataset. (a) Rotated input image (b) Baseline result (HRNet-W32) [21] (c) Proposed method result. Note that these figures depict the results for every 90 degrees, but our experiments were conducted by rotating the images every 15 degrees.

D. EVALUATION PROTOCOL

The proposed method followed the general keypoint detection protocol; however, we designed experiments to measure...
the robustness of keypoint detection against angular changes. The performance was evaluated by rotating the input image through 15 degrees at a time, from 0 to 360 degrees, via rotation transformations and by rotating the ground truth correspondingly, as depicted in Figure 3. For the comparison, we applied the proposed method to various backbone networks of SimpleBaseline [20] and HRNet [21], which are state-of-the-art methods for human pose estimation.

For training and validation data, we used the following splits for each dataset.

- MPII human pose estimation dataset [24]: For over 40K instances, 28,821 instances were used for the training set and 11,701 instances were used as the test set.
- COCO keypoint detection dataset [19]: For over 150K people, we used the official “train2017” (149,813 people) and “val2017” (6,352 people) splits for the training and validation sets, respectively.

V. EXPERIMENTS

We verify our arguments using various experiments whose setups are described in Section IV. Figure 4 depicts a set of experiments designed for the evaluation of the proposed method and a summary of the experimental figures and tables.

- Quantitative experiments were conducted based on the MPII human pose estimation dataset [24] and COCO keypoint detection dataset [19].
  - Experiments on two state-of-the-art methods: rotational robustness test (Section V-A) and rotational data augmentation effects (Section V-B)
  - Experiments on the proposed method: rotational robustness test (Section V-C.1), ablation study (Section V-C.2), statistical analysis of the reported performances (Section V-C.3), and comparison results of many existing methods (Section V-C.4)
- Based on a real-world surveillance dataset [25], [33], ICVL dataset [26], and IASLAB Fallen Person dataset [27], we present qualitative results to verify the robustness of the proposed network against rotation (Section V-D).

A. ANALYSIS OF ROTATIONAL ROBUSTNESS FOR THE TWO BASELINE METHODS

We analyze the changes in the performances of the existing methods corresponding to each rotation of the input image. The circular graphs in Figures 5 and 6 illustrate the performance changes for existing methods [20], [21] with respect to the MPII and COCO datasets. The distance between the center and each angle represents the performance when the image is rotated clockwise by that angle. The value corresponding to 0 degrees, located at the 12 o’clock position, is considered as the performance of the original human pose estimation, and the performance difference is observed to depend on each method and the backbone network used. In an ideal scenario, where the performance is independent of the angle of rotation, the graph would be a perfect circle. However, as is apparent from Figure 5(a), the performances of all existing methods drop from 80 to 70 at 90 degrees and then rapidly drop further to 40 when rotated by 180 degrees. In addition to Figure 5(a), which depicts the overall performance, the methods have the same tendency while estimating the five joints individually, as depicted in Figure 5(b)–(f). In each case, the performance decreases rapidly with the rotation of the input image.

As recorded in Figure 6, the COCO dataset was more difficult to handle than the MPII dataset, and the methods exhibited much larger performance drops in its case. Modern networks, such as HRNet-W48, produced improved performances compared to other backbones; however, they were vulnerable to inaccuracies arising from rotational changes. Our experimental results demonstrate that conventional methods are biased with respect to the shape of human appearance (e.g., with the head at the top). Owing to this tendency, conventional methods are weak when the posture of the person is an uncommon one, such as in dynamic sports or in surveillance scenes caused by the camera angle.

B. ANALYSIS OF THE EFFECTS OF A LARGE AMOUNT OF ROTATION AUGMENTATION

Data augmentation is the most intuitive approach to create a network that is robust with respect to rotational changes. Existing algorithms apply data augmentation using random rotations in the range of [−30, 30] degrees for MPII datasets and random rotations in the range of [−45, 45] degrees for COCO datasets. To measure the robustness of the existing networks utilizing data augmentation, we applied data augmentation with random rotations in the range of [−90, 90] degrees to each dataset. To measure the robustness of the proposed method, the same random rotation augmentation explained in Section IV-A was used. This setup allowed us to compare the robustness of the proposed method with that of the simple data augmentation method.

Figure 7 records improvements in the performance and effectiveness of the proposed method and the rotation augmentation method on the MPII dataset. In Figure 7, Aug90 denotes the rotation augmentation method, and Proposed denotes the proposed method. The x-axis of the graph represents angles ranging from 0 degrees to 360 degrees, and
FIGURE 5. Rotational robustness test for MPII human pose estimation. The distance between the center of the circle and each angle represents the performance when the input image is rotated clockwise by that angle. The twelve o’clock position was set to 0 degrees and displayed clockwise. (a) depicts the average performance and (b)–(f) depict each performance corresponding to the elbow, wrist, hip, knee, and ankle joints.

FIGURE 6. Rotational robustness test for COCO keypoint detection. The distance between the center of the circle and each angle represents the performance when the input image is rotated clockwise by that angle. The twelve o’clock position was set to 0 degrees and displayed clockwise. (a) depicts the average precision and (b)–(c) depict the average precision for medium-sized persons and large-sized persons, respectively.

As depicted in Figure 7, both SimpleBaseline and HRNet baseline models exhibit a sharp drop in performance with changes in the angle of rotation, and the rotation data augmentation yields significantly better performance as is apparent from the differences between the blue and green loci.
FIGURE 7. Performance comparison of the rotation augmentation and the proposed method in the MPII dataset.

However, even though the proposed method uses the same angular data augmentation procedure, its performance (red locus) is better than that of the rotation augmentation method (green locus).

In the COCO dataset, which is more difficult to handle than the MPII dataset, the improvement in performance between the proposed method and simple rotation augmentation is more readily apparent. Figure 8 records the improvements in performance and the effectiveness of the proposed method and rotation augmentation on the COCO dataset. The performances of the conventional methods deteriorate rapidly with an increase in the angle of rotation, as depicted by the blue locus, and rotation data augmentation performed slightly better, as is apparent from the difference between the blue and green loci. However, the proposed method performs significantly better than simple data augmentation by combining the two-path outputs from the shared parameters using the Siamese network.

C. ANALYSIS OF ROTATIONAL ROBUSTNESS OF THE PROPOSED METHOD
1) EVALUATION
The proposed method is applied to various human pose estimation backbone networks [20], [21] to analyze the improvement in the performance. Figure 9 records the improvements in the rotational robustness of the proposed method as a circle graph. On comparing Figure 9 with Figures 5 and 6 in Section V-A, the rotational robustness is observed to have greatly improved. The proposed method is not limited in its application to the structure of backbone networks solely, and it improves the rotational robustness of both the SimpleBaseline and HRNet human pose estimation methods. In particular, for the COCO dataset, the performances of conventional methods, as depicted in Figure 6, deteriorated to nearly zero; however, the proposed method made the network robust against rotation, as depicted in Figure 9(b).

2) ABLATION STUDY
Ablation study is conducted to analyze the performances of each of the two paths (original and degree 90) in the proposed method. As illustrated in the network structure presented in Figure 2, the proposed method shares its parameters with the backbone network and generates each feature map of the original image and rotated image by 90 degrees, respectively. Following that, it concatenates the two feature maps in the channel direction and passes them to a combined module to estimate the final keypoint.

Figures 10 and 11 present the performances as measured on the MPII and COCO datasets, respectively. The blue locus represents the performance of the case in which the original input is used as Path 1, and the green locus represents the performance of the case in which the input image that has been rotated by 90 degrees is used as Path 2. As depicted in Figures 7 and 8, the human pose estimation network exhibits the poorest performance when the input image is rotated by 180 degrees. Similarly, as depicted in Figures 10 and 11, Path 1, represented here by the blue locus, also exhibits the poorest result in the case of rotation by 180 degrees. As Path 2 already uses the image rotated by 90 degrees as input, its performance at 90 degree rotation is the same as that of the original rotated by 180 degrees. Thus, its performance is poorest for rotation by 90 degrees.

FIGURE 8. Performance comparison of the rotation augmentation and the proposed method with respect to the COCO dataset.
The performance of the proposed network that combines these two feature maps is always better than the individual performance of each path, much like a voting scheme.

Figure 10(c)–(d) and Figure 11(d)–(f) are circular graphs that present the robustness of the algorithm. Path 1 performs slightly worse at 180 degrees and Path 2 at 90 degrees, but the final combined model performs better than each of the individual components, as represented by its circle graph.

In addition, the proposed network uses images rotated by 0 degrees and 90 degrees simultaneously to train the network weights. Thus, the backbone is trained to be a more robust feature extractor, and it avoids the performance degradation of the conventional method, as depicted in Figure 5 – both Path 1 and Path 2 approximate the circle except at around 90 degrees.

3) STATISTICAL ANALYSIS FOR THE ROTATIONAL CHANGE TEST

Let a result of the human pose estimation performance from one rotation angle be a sample of the pose estimation performance. Then, we can fit the experimental results for all rotational angles to a probability model for the test results. We estimated the mean and standard deviations by fitting a normal distribution to the samples from the rotational changes. Figures 12 and 13 depict both test samples and the estimated probability density functions (pdf) with respect to the expected performance x-axis, assuming that the rotation angle of the input image is random. As with the purple graph, the conventional method shows very large deviations and a low average performance caused by the rotation of the input image. Using data augmentation shows a slightly improved performance as shown in the blue graph; however, the performance remains unsatisfactory. In contrast, the proposed method, represented by the red graph, exhibits the best mean performance and the lowest standard deviation. In particular, only each component path (degree 0 and degree 90) of the proposed method shows similar performance in the MPII dataset and better performance in the COCO dataset.
FIGURE 11. Path 1 (original image), Path 2 (image rotated by 90 degrees), and the combined result (Proposed) in the COCO dataset.

FIGURE 12. Samples from rotational change test and the estimated probability density function by fitting a normal distribution on the MPII dataset.

The conventional method has a 95% chance of achieving a performance between 47.635 and 63.716. In other words, the proposed method shows that the average performance is significantly increased, and the standard deviation is significantly lowered for rotational change. This statistical analysis shows that the proposed method is a robust pose estimator that considerably improves the conventional pose estimator that is vulnerable to rotational changes.

4) PERFORMANCE OF ORIGINAL HUMAN POSE ESTIMATION PROBLEM

Tables 3 and 4 present the performance of original human pose estimation problem (i.e., rotation by 0 degree) on the MPII and COCO datasets, respectively. In general, robustness is directly proportional to the quantity of data augmentation applied. However, the overuse of the technique reduces the generality of the feature and might cause slight degradation in performance. Although performances differ depending on the type of the backbone network and the dataset used, the overall performance of rotation augmentation deteriorates, as illustrated in Table 4. In contrast, the proposed method exhibits lesser performance degradation than data augmentation on the COCO dataset and better performance than baseline methods on the MPII dataset.

compared to the data augmentation method. Thus, each path constituting the proposed method becomes a robust feature extractor via two-path training.

Tables 1 and 2 show the estimated parameters (mean and standard deviation, respectively) of a normal distribution. The tables also show the 95% confidence intervals [34], [35] of the estimated parameters. When an input with arbitrary rotational change is given to the HRNet-W48 pose estimator on the COCO dataset, the proposed method has a 95% probability that the performance is between 69.762 and 73.169.
Table 4 also records the network parameters of the baseline methods and the proposed method. The parameters used in the proposed method are increased slightly because of the convolutional layers used to combine two path inputs; however, the increment is negligible because the parameters are shared with the backbone networks. In the case of the SimpleBaseline method based on ResNet-50, the number of network parameters is increased by 1.18M. In the HRNet structure, the number of network parameters is increased from 0.02M (HRNet-W32) to 0.04M (HRNet-W48) because of the small number of parameters of the last combined module. Owing to the structure and learning procedure of the proposed method, rotational robustness can be greatly improved by marginally increasing the number of network parameters.

**D. QUALITATIVE RESULT**

Real-world images often involve people in rotated postures as mentioned in the introduction. For example, the image of a person dumping garbage captured by a surveillance camera exhibits rotation caused by the bending pose.
Figures 14 and 15 qualitatively visualize the human pose estimation results of the baseline method (HRNet-W32 trained on COCO dataset) and the proposed method on the surveillance action dataset [25], [26]. Figure 14 presents the results of human pose estimation in the case in which the image is rotated according to the evaluation protocol on the garbage dumping action dataset [25], [33]. As depicted in Figure 14, rotational robustness is improved on the actual surveillance dataset as well as the MPII and COCO datasets used in the quantitative evaluation.

Figure 15 presents the results of human pose estimation on several images labeled “garbage dumping action” in the ICVL dataset [26]. The baseline method experiences difficulties in locating the head and the torso on this dataset. Human postures are significantly different from those in the conventional dataset because of the incongruity between the camera’s installation angle and the actual posture of the dumping action. In contrast, as the variation in data from real CCTV images is similar to that caused by rotation, the proposed method improves the robustness against rotation.
even if the head or torso is not placed in the usual position.

Further, we evaluated the human pose estimation on the IASLAB Fallen Person dataset [27] captured by a home camera and a mobile robot. The head of a reclining person is oriented upside-down or sideways, which is different from the configuration in conventional human pose estimation datasets. Thus, as recorded in the first row of Figure 16, the existing method experiences difficulty in locating the posture of reclining person. In contrast, the proposed method estimated the posture of reclining persons more accurately because the associated inputs are similar to images obtained via rotation transformation.

In other words, although human posture may not always have the usual form due to the camera view and appearances because of specific actions, the proposed method can detect human joints more accurately and consistently than the baseline method.

| Method         | Backbone     | +Flip | Input Size | #Params | AP  | AP$^{50}$ | AP$^{55}$ | AP$^{75}$ | AP$^{M}$ | AP$^{L}$ | AR  |
|----------------|--------------|-------|------------|---------|-----|-----------|-----------|-----------|----------|----------|-----|
| 8-stage Hourglass [6] | ResNet-50    | Y     | 256 x 192  | 25.1M   | 66.9| -         | -         | -         | -         | -        | -   |
| CPN [12]       | ResNet-50    | Y     | 256 x 192  | 27.0M   | 68.6| -         | -         | -         | -         | -        | -   |
| CPN + OHKM [12] | ResNet-50    | Y     | 256 x 192  | 27.0M   | 69.4| -         | -         | -         | -         | -        | -   |
| SimpleBaseline | ResNet-50    | Y     | 256 x 192  | 34.0M   | 70.4| 88.6      | 78.3      | 67.1      | 77.2      | 76.3     | -   |
| SimpleBaseline | ResNet-101   | Y     | 256 x 192  | 53.0M   | 71.4| 89.3      | 79.3      | 68.1      | 78.1      | 77.1     | -   |
| SimpleBaseline | ResNet-152   | Y     | 256 x 192  | 68.6M   | 72.0| 89.3      | 79.8      | 68.7      | 78.9      | 77.8     | -   |
| SimpleBaseline | ResNet-152   | Y     | 384 x 288  | 68.6M   | 74.3| 89.6      | 81.1      | 70.5      | 79.7      | 79.4     | -   |
| HRNet-W32      | HRNet-W32    | Y     | 256 x 192  | 28.53M  | 74.4| 90.5      | 81.9      | 70.8      | 81.0      | 80.9     | -   |
| HRNet-W32      | HRNet-W32    | Y     | 384 x 288  | 28.53M  | 75.8| 90.6      | 82.7      | 71.9      | 82.8      | 81.0     | -   |
| HRNet-W48      | HRNet-W48    | Y     | 256 x 192  | 63.60M  | 73.1| 90.9      | 82.2      | 71.5      | 81.8      | 80.4     | -   |
| HRNet-W48      | HRNet-W48    | Y     | 384 x 288  | 63.60M  | 76.3| 90.8      | 82.9      | 72.3      | 83.4      | 81.2     | -   |
| SimpleBaseline | ResNet-50    | N     | 256 x 256  | 34.00M  | 71.2| 91.4      | 78.1      | 68.1      | 76.1      | 74.4     | -   |
| SimpleBaseline | ResNet-50    | N     | 256 x 256  | 34.00M  | 70.8| 91.5      | 78.2      | 67.2      | 75.6      | 73.9     | -   |
| SimpleBaseline | ResNet-50    | N     | 256 x 256  | 35.18M  | 71.1| 91.4      | 79.2      | 67.5      | 75.6      | 76.4     | -   |
| SimpleBaseline | ResNet-50    | Y     | 256 x 256  | 34.00M  | 72.5| 91.6      | 80.3      | 69.3      | 77.1      | 75.5     | -   |
| SimpleBaseline | ResNet-50    | Y     | 256 x 256  | 34.00M  | 71.7| 91.5      | 79.4      | 68.6      | 76.6      | 74.9     | -   |
| SimpleBaseline | ResNet-50    | Y     | 256 x 256  | 35.18M  | 72.2| 92.5      | 80.4      | 68.8      | 77.3      | 75.3     | -   |
| HRNet-W32      | HRNet-W32    | N     | 256 x 256  | 28.53M  | 75.1| 92.6      | 82.6      | 72.1      | 80.7      | 78.1     | -   |
| HRNet-W32     | HRNet-W32    | N     | 384 x 288  | 28.53M  | 74.8| 92.4      | 82.5      | 71.7      | 79.5      | 77.8     | -   |
| HRNet-W32     | HRNet-W32    | N     | 256 x 256  | 28.55M  | 74.9| 93.5      | 82.6      | 71.7      | 80.7      | 78.0     | -   |
| HRNet-W32     | HRNet-W32    | N     | 256 x 256  | 28.55M  | 76.4| 93.6      | 83.8      | 73.3      | 81.2      | 79.1     | -   |
| HRNet-W32     | HRNet-W32    | N     | 256 x 256  | 28.55M  | 75.4| 93.6      | 83.6      | 72.2      | 80.4      | 78.5     | -   |
| HRNet-W48     | HRNet-W48    | N     | 256 x 256  | 63.60M  | 76.5| 93.5      | 83.6      | 72.9      | 81.5      | 79.1     | -   |
| HRNet-W48     | HRNet-W48    | N     | 256 x 256  | 63.60M  | 76.4| 93.5      | 83.6      | 73.3      | 81.3      | 79.4     | -   |
| HRNet-W48     | HRNet-W48    | N     | 256 x 256  | 63.64M  | 76.5| 93.5      | 84.5      | 72.6      | 82.1      | 79.3     | -   |
| HRNet-W48     | HRNet-W48    | N     | 256 x 256  | 63.60M  | 77.4| 93.6      | 84.8      | 74.1      | 82.3      | 80.1     | -   |
| HRNet-W48     | HRNet-W48    | N     | 256 x 256  | 63.60M  | 77.7| 93.6      | 84.7      | 74.7      | 82.1      | 80.4     | -   |
| HRNet-W48     | HRNet-W48    | N     | 256 x 256  | 63.64M  | 77.7| 93.6      | 84.8      | 74.4      | 82.6      | 80.4     | -   |

**FIGURE 16.** Examples of results of the conventional human pose estimation and the proposed human pose estimation on the IASLAB Fallen Person dataset [27].

| TABLE 4. | Comparisons on the COCO validation set. All tests are based on the ground truth of the human bounding box. +Flip indicates the use of both the original image and the horizontally flipped image. OHKM is online hard keypoint mining [12]. |
VI. CONCLUSION
In this study, we analyzed the robustness of human pose estimation with respect to rotational changes through large angles and proposed a novel method to improve rotational robustness. Existing methods perform well for typical human postures on the conventional human posture dataset, but the performance deteriorates drastically in images that appear to be rotated due to incongruence with the installation angle of real-world cameras. To address this problem, we designed a robust human pose estimator based on a self-supervised method that improves the learning of rotational changes on images. The proposed method is combined with a state-of-the-art human pose estimation network, and it is found to greatly improve rotational robustness. We conducted extensive analysis and experiments to verify the rotational robustness of the proposed method and compared the results with those of conventional methods and data-augmentation-based methods on the MPII and COCO datasets.

REFERENCES
[1] L. Ke, M.-C. Chang, H. Qi, and S. Lyu, “Multi-scale structure-aware network for human pose estimation,” in Proc. Eur. Conf. Comput. Vis., 2018, pp. 713–728.
[2] D. C. Luvizon, D. Picard, and H. Tabia, “2D/3D pose estimation and action recognition using multitask deep learning,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 5137–5146.
[3] S. Yan, Y. Xiong, and D. Lin, “Spatial temporal graph convolutional networks for skeleton-based action recognition,” in Proc. AAAI Conf. Artif. Intell., Jan. 2018, pp. 7444–7452.
[4] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black, “Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image,” in Proc. Eur. Conf. Comput. Vis., 2016, pp. 561–578.
[5] S. Xia, L. Gao, Y.-K. Lay, M.-Z. Yuan, and J. Chai, “A survey on human performance capture and animation,” J. Comput. Sci. Technol., vol. 32, no. 3, pp. 536–554, May 2017.
[6] A. Newell, K. Yang, and J. Deng, “Stacked hourglass networks for human pose estimation,” in Proc. Eur. Conf. Comput. Vis. Comput. Vis., 2016, pp. 483–499.
[7] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, “Realtime multi-person 2D pose estimation using part affinity fields,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 7291–7299.
[8] W. Yang, S. Li, W. Ouyang, H. Li, and X. Wang, “Learning feature pyramids for human pose estimation,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 1281–1290.
[9] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, “RMPE: Regional multi-person pose estimation,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 2334–2343.
[10] S. Gidaris, P. Singh, and N. Komodakis, “Unsupervised representation learning by predicting image rotations,” 2018, arXiv:1803.07728. [Online]. Available: http://arxiv.org/abs/1803.07728
[11] T. Seki, “Pose proposal networks,” in Proc. Eur. Conf. Comput. Vis., 2018, pp. 342–357.
[12] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun, “Cascaded pyramid network for multi-person pose estimation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 7103–7112.
[13] G. Moon, J. Y. Chang, and K. M. Lee, “PoseFix: Model-agnostic general human pose refinement network,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 7773–7781.
[14] Y. Yang and D. Ramanan, “Articulated pose estimation with flexible mixtures-of-parts,” in Proc. CVPR, Jun. 2011, pp. 1385–1392.
[15] M. Eichner, M. Marin-Jimenez, A. Zisserman, and V. Ferrari, “2D articulated human pose estimation and retrieval in (almost) unconstrained still images,” Int. J. Comput. Vis., vol. 99, no. 2, pp. 190–214, Sep. 2012.
[16] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2009, pp. 248–255.
[17] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 770–778.
[18] A. Toshev and C. Szegedy, “DeepPose: Human pose estimation via deep neural networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2014, pp. 1653–1660.
[19] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft COCO: Common objects in context,” in Proc. Eur. Conf. Comput. Vis., 2014, pp. 740–755.
[20] B. Xiao, H. Wu, and Y. Wei, “Simple baselines for human pose estimation and tracking,” in Proc. Eur. Conf. Comput. Vis., 2018, pp. 466–481.
[21] K. Sun, B. Xiao, D. Liu, and J. Wang, “Deep high-resolution representation learning for human pose estimation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 5693–5703.
[22] L. Zhang, G.-J. Qi, L. Wang, and J. Luo, “AET vs. AED: Unsupervised representation learning by auto-encoding transformations rather than data,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 2547–2555.
[23] Z. Feng, C. Xu, and D. Tao, “Self-supervised representation learning by rotation feature decoupling,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 10364–10374.
[24] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, “2D human pose estimation: New benchmark and state of the art analysis,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2014, pp. 3686–3693.
[25] K. Yun, Y. Kwon, S. Oh, J. Moon, and J. Park, “Vision-based garbage dumping action detection for real-world surveillance platform,” ETRI J., vol. 41, no. 4, pp. 494–505, Aug. 2019.
[26] C.-B. Jin, S. Li, and H. Kim, “Real-time action detection in video surveillance using sub-action descriptor with multi-CNN,” 2017, arXiv:1710.03383. [Online]. Available: http://arxiv.org/abs/1710.03383.
[27] M. Antonello, M. Carraro, M. Pierobon, and E. Menegatti, “Fast and robust detection of fallen people from a mobile robot,” in Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS), Sep. 2017, pp. 4159–4166.
[28] CAVIAR: Context Aware Vision Using Image-Based Active Recognition. Accessed: Nov. 20, 2019. [Online]. Available: http://homepages.inf.ed.ac.uk/rbf/CAVIAR/caviar.htm
[29] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in Proc. Adv. Neural Inf. Process. Syst., 2012, pp. 1097–1105.
[30] M. Jaderberg, K. Simonyan, and A. Zisserman, “Spatial transformer networks,” in Proc. Adv. Neural Inf. Process. Syst., 2015, pp. 2017–2025.
[31] J. Foley, A. van Dam, S. Feiner, J. Hughes, and R. Phillips, Introduction to Computer Graphics. Reading, MA, USA: Addison-Wesley, 1994.
[32] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” 2014, arXiv:1412.6980. [Online]. Available: http://arxiv.org/abs/1412.6980.
[33] S. Kim, K. Yun, J. Park, and J. Y. Choi, “Skeleton-based action recognition of people handling objects,” in Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV), Jan. 2019, pp. 61–70.
[34] A. W. Bowman and A. Azzalini, Applied Smoothing Techniques for Data Analysis: The Kernel Approach With S-Plus Illustrations. New York, NY, USA: Oxford Univ. Press, 1997.
[35] D. J. Sheskin, Handbook parametric nonparametric statistical procedures. Boca Raton, FL, USA: CRC Press, 2003.
[36] W. Tang, P. Yu, and Y. Wu, “Deeply learned compositional models for human pose estimation,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2018, pp. 190–206.

KIMIN YUN received the B.S. and Ph.D. degrees from the Department of Electrical and Computer Engineering, Seoul National University, Seoul, South Korea, in 2010 and 2017, respectively. Since 2017, he has been working with the Artificial Intelligence Research Laboratory, Visual Intelligence Research Institute, Daejeon, South Korea. His current research interests include machine learning, computer vision, visual event analysis, moving object detection, and video analysis.
JONGYOUL PARK received the B.S. degree in computer engineering from Chungnam National University, Daejeon, South Korea, in 1996, and the M.S. and Ph.D. degrees in information and communication engineering from the Gwangju Institute of Science and Technology, South Korea, in 1999 and 2004, respectively. From 2001 to 2002, he was a Visiting Researcher with the School of Computing, The University of Utah, UT, USA. Since 2004, he has been working with the Artificial Intelligence Laboratory, Visual Intelligence Research Section, Electronics and Telecommunications Research Institute, Daejeon. His research interests include visual intelligence, machine learning, and behavior and scene.

JUNGCHAN CHO (Member, IEEE) received the B.S. degree from the School of Electrical and Electronics Engineering, Chung-Ang University, Seoul, South Korea, in 2010, and the Ph.D. degree from the Department of Electrical and Computer Engineering, Seoul National University, Seoul, in 2016. From 2016 to 2019, he was a Senior Software Engineer with Samsung Electronics. He is currently an Assistant Professor with the Department of Software, Gachon University, Seongnam, South Korea. His research interests include deep learning, computer vision, and machine learning.