

Multi-resolution Fusion Network for Human Pose Estimation in Low-resolution Images

Boeun Kim, YeonSeung Choo, Hea In Jeong, Chung-II Kim, Saim Shin and Jungho Kim*

Artificial Intelligence Research Center, Korea Electronics Technology Institute, Korea

[e-mail: kbe36@keti.re.kr, piksal@keti.re.kr, heain@keti.re.kr, cilkim1@keti.re.kr,

sishin@keti.re.kr, jhkim77@keti.re.kr]

*Corresponding author: Jungho Kim

*Received November 25, 2021; revised January 29, 2022; accepted March 1, 2022;
published July 31, 2022*

Abstract

2D human pose estimation still faces difficulty in low-resolution images. Most existing top-down approaches scale up the target human bounding box images to the large size and insert the scaled image into the network. Due to up-sampling, artifacts occur in the low-resolution target images, and the degraded images adversely affect the accurate estimation of the joint positions. To address this issue, we propose a multi-resolution input feature fusion network for human pose estimation. Specifically, the bounding box image of the target human is rescaled to multiple input images of various sizes, and the features extracted from the multiple images are fused in the network. Moreover, we introduce a guiding channel which induces the multi-resolution input features to alternatively affect the network according to the resolution of the target image. We conduct experiments on MS COCO dataset which is a representative dataset for 2D human pose estimation, where our method achieves superior performance compared to the strong baseline HRNet and the previous state-of-the-art methods.

Keywords: Human keypoint detection, Human pose estimation, Low-resolution image, Small person pose estimation, 2D pose estimation

1. Introduction

2D human pose estimation determines the pixel locations of body joints in a given single image.

Estimating human pose has received considerable attention in the field of computer vision, and has made great progress with the introduction of deep learning [1-20]. Pose estimation plays an important role in various human-computer interaction tasks, such as surveillance systems and autonomous driving. In the applications, precisely estimated joint positions are leveraged as features of action and gesture recognition.

Many recent studies show the good performance in high-resolution images, however, as the resolution decreases, the performance drastically decreases as well [4-9]. The same problem happens when the size of the person in the image becomes smaller because the person is far from a camera. However, for practical industrial applications, it is important to estimate the pose of a distant person due to the limited number of cameras in the field. In addition, in order to process images captured by various camera systems, a model that robustly operates in low-resolution images is essential. In this paper, we propose a model that accurately estimates 2D human pose in low-resolution human images while maintaining the performance for images for various sizes.

2D human pose estimation methods are classified into two categories, bottom-up and top-down approaches. The bottom-up methods first obtain joint candidates for the input image, and then estimate the pose of each person by analyzing the correlation between the obtained joints [4, 5]. On the other hand, the top-down methods first detect the person in the image, and then the positions of joints are estimated from the region of the detected person [6-9, 18-19]. In general, the bottom-up methods are faster than the top-down methods when several persons exist in one image. Instead, the top-down methods generally show higher accuracy than the bottom-up methods. Our proposed model is based on the top-down approach to predict joint positions of the persons.

There are a few existing studies on 2D pose estimation focusing on low-resolution images. The method in [13] is a firstly proposed pose estimation model for low-resolution images in which tiny people dataset was generated to validate their model, but the dataset is currently not available. In the field of 3D human pose estimation, the method in [14] attempts to estimate accurate 3D joint positions from low-resolution images. This method has improved the performance by jointly training several networks of the same structure whose inputs are images with different resolutions. The training scheme in [14] improved the performance of 3D pose estimation, however, the memory consumption increases depending on the number of network parameters trained simultaneously, making it difficult to apply it to models with a large number of parameters. Recently, a method of scaling up the image using a deep learning-based super-resolution model, and then applying the generated high-resolution image to the Hourglass pose estimation network [6] has been proposed [15]. Although this method greatly improved the performance for 32×24 human images, it has a weakness that a single model cannot cope with various image resolutions in that the super-resolution model operates at a fixed magnification. In addition, it is difficult to be used in real-time applications due to the computational burden caused by the super-resolution module.

In this paper, we propose a model that improves 2D joint estimation accuracy in low-resolution images with little increase in the computational cost on the basis of the structure of HRNet [9], which is a strong baseline among the current top-down methods. Previous methods scale the region of the person to the fixed-sized image to leverage it as an input to the network [6-9]. In those methods, excessive artifacts may occur in a low-resolution image scaled up by

a large ratio, which consequently causes errors in pose estimation. On the other hand, scaling down a high-resolution image brings a loss of information. To overcome this problem, we scale the original bounding box for the person image to multiple resolutions and leverage all of the scaled images as inputs in the following pose estimation network. In addition, we introduce the guiding channel to let the network know which input image is closest to the original image size. This allows a single network to work robustly on the images with various resolutions. The proposed method improves the performance of 2D pose estimation by adding a few light convolution layers without computationally expensive modules, such as a super-resolution module.

To summarize, we make the following contributions in this work:

- We propose a network that effectively estimates 2D human poses on images of various resolutions, especially low-resolution images, through the multi-scale input feature fusion and a guiding channel.
- We find the optimal model among various network structures and the network input image sizes through experiments.
- On a representative dataset for 2D human pose estimation, MS COCO [21], the proposed model achieves superior performance compared to the strong baseline HRNet [9] and previous state-of-the-art methods. Qualitative comparison against the baseline model is provided and it is verified that our model works appropriately on the target surveillance system.

2. Related Work

2.1 2D pose estimation

Existing methods before the advent of deep learning have used hand-crafted features such as Histogram of Oriented Gradients (HOG) and Local Binary Pattern (LBP) to estimate human features [1-3]. Recently, pose estimation methods using deep learning have shown the great performance [4-20]. 2D human pose estimation has been developed by two approaches, top-down and bottom-up approaches. The top-down approach first detects the regions of the persons from the image and then estimates the joint positions of the person within the detected region. The bottom-up approach first detects human joint candidates in the given image, and it connects the detected joints to the human skeleton structure by considering the correlation between the joints. The bottom-up methods are generally faster than the top-down methods. However, the estimation accuracy of the bottom-up approach is lower than that of the top-down approach because it is difficult to connect joint candidates belonging to one person without region information of the human especially when several persons with different scales exist in the image.

One of the representative bottom-up methods is OpenPose [4] which finds joint candidates in the image by generating a part confidence map. After then, joints belonging to each person are grouped by estimating the relationship between joints from the part affinity field. Recently, a method for improving performance by combining a top-down method with a bottom-up method has been proposed [5, 16, 17]. In the method in [5], the joint position estimation and the person detection were simultaneously performed to find the accurate joint associations by resolving the performance degradation when there exist multiple persons with different scales, which is the fundamental problem of the bottom-up methods

Current state-of-the-art top-down methods improves their performance by introducing a high-to-low and low-to-high framework that scales down and then scales up the input image while

passing through the network [6-9, 18, 19]. Hourglass [6] improved the pose estimation accuracy by stacking autoencoder networks including symmetrical layers and residual connections. In addition, SimpleBaseline [7] is a network that introduces a high-to-low process composed of multiple deep layers and relatively light low-to-high process. Cascaded Pyramid Network (CPN) [8] is composed of a cascade process that localizes evident joints in the pyramid network and then finds difficult joints in the separated refinement network. HRNet [9] introduce a multi-scale fusion network that fuses high-resolution and low-resolution features which are generated in the middle layers in the network. This method greatly improved the performance of 2D pose estimation, and many recent methods leveraged this framework as a baseline [10]. In this study, we propose a model that maintains the average accuracy of pose estimation in all resolutions of the input images while showing superior performance especially for low-resolution images on the basis of the structure of HRNet [9].

2.2 Pose estimation for low-resolution images

2D pose estimation methods basically aim to be robust to different resolutions of images. In general, the performance for the low-resolution image is significantly lower than that of the high-resolution image. In order to apply the model to real world applications, pose estimation of the person far away from a camera is important due to the limited number of cameras in the field. However, there are a few studies focusing on pose estimation for small persons or low-resolution images. Therefore, we first review object detection and face detection in low-resolution images, which is a similar research field, and then introduce studies related to 2D and 3D pose estimation.

Several deep learning-based methods for small object detection have been proposed. In the methods, low-resolution images are scaled up and the detection process is performed in the generated high-resolution images [11-12]. The method in [11] first detects face candidates and upscale them using a generator network consisting of several convolution and deconvolution layers. Then a discriminator network classifies face or non-face images. The generator and the discriminator are trained in an adversarial manner. The method in [12] leverages the super-resolution module to scale up the image size of the candidate object estimated by the object detector, and then performs the precise object detection process again with the high-resolution images.

The method of [13] is the first 2D pose estimation model focusing on small people in the image. The probabilistic model was introduced for modeling the ambiguity that arises from estimating human poses in small images. This model estimates posterior probability maps for all joints to regress the joint positions, and each probability map consists of a Gaussian mixture model for semi-dense subpixels. The 3D pose estimation model which is robust for low-resolution images was first proposed in [14]. Input images with different resolutions are trained on the different networks with same structure, and the resolution-dependent parameters of each network were adaptively integrated to induce a robust model for images of various sizes. In addition, the performance of 3D pose estimation using the low-resolution image was improved by allowing the results of the network assigned high-resolution images to guide the learning of the network assigned low-resolution images. This scheme is difficult to apply to a model with a large number of parameters, because the number of parameters doubles according to the number of resolutions used. The method of [15] generates high-resolution images by inserting the pre-detected region of interest into the super resolution module, and then performs pose estimation with the scaled up image using the Hourglass [6] model. Because the super resolution module is trained with a fixed magnification, there is a limitation that only the input images of the fixed resolution range can be accepted for the whole network.

Furthermore, as the super resolution network is added, the number of parameters increases, slowing down for pose estimation and making it difficult to be used in real-time systems.

In this study, we propose a 2D pose estimation model that uses multi-resolution images as inputs to the network to be robust on both low-resolution and high-resolution images. By adding only a few stem layers, the proposed model outperforms HRNet [9] for the low-resolution images while maintaining the average performance of various sized images.

3. Method

3.1 HRNet Model Architecture

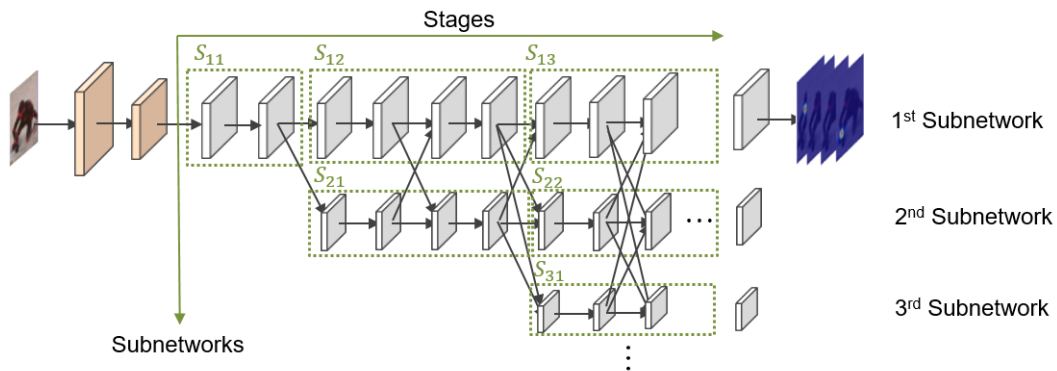


Fig. 1. The framework of HRNet [9] with notations.

In contrast to the serial connection of layers in the previous 2D pose estimation networks, HRNet [9] achieved high performance by configuring the layers both in parallel and in series. In HRNet [9], several high-to-low subnetworks are connected in parallel, where several stages are serially connected in each high-to-low subnetwork. If the k -th stage of the n -th subnetwork is denoted by S_{nk} . The whole framework is illustrated in Fig. 1.

Each stage involves a multi-scale fusion process that aggregates features from different subnetworks. The input features of S_{nk} aggregates the output features of $S_{1(n+k-2)}$, $S_{2(n+k-3)}$, \dots , $S_{(n+k-2)1}$. For example, the features from S_{13} , S_{22} , and S_{31} are combined and inserted into stage S_{23} . Strided convolution or up-sampling is used to match the size of feature maps to be integrated. To generate an input feature of S_{nk} , we reduce the size of the output feature $S_{(n-1)k}$ by half by strided convolution. The size of the output feature of stage $S_{(n+1)(k-2)}$ is doubled through nearest neighbor up-sampling and 1×1 convolution. From $S_{n(k-1)}$, the feature map passes through the convolution layer without change in size. The input feature of S_{nk} is generated by summing those feature maps with the same size.

The heat maps to regress the joint position are obtained from the output feature of the first subnetwork. The ground truth heat maps are generated by applying 2D Gaussian with a standard deviation of 1 pixel from the ground truth pixel locations of each joint. The output representation is trained to approximate the ground truth heat maps using the mean squared error loss.

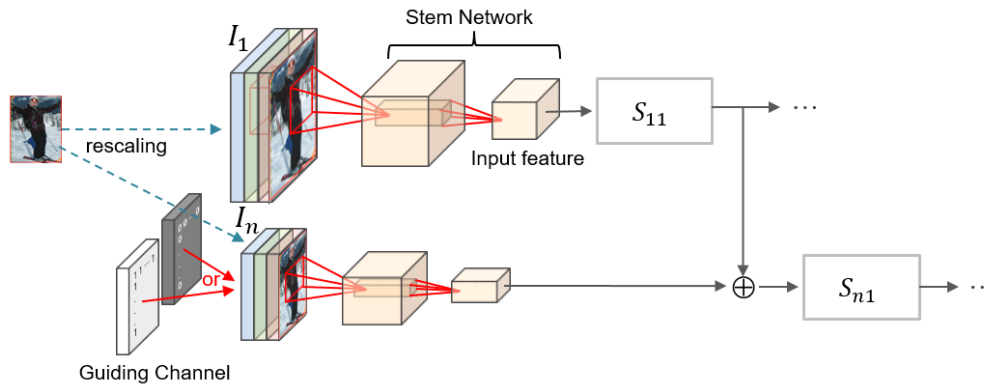


Fig. 2. Multi-resolution inputs and a guiding channel. The original target image is rescaled to various resolutions. At $N > 2$, a guiding channel filled with binary values is added as the fourth channel of the image. The features of the multi-resolution input images are extracted through stem networks and used for estimating joint positions. The features of the n -th subnetwork are added to the downsampled features from the $(n-1)$ -th subnetwork and inserted into the network

3.2 Multi-resolution Input Feature Fusion Network

3.2.1 Multi-resolution Input Feature

We propose a method to leverage multi-resolution input images for the estimation of 2D joint positions, making the framework robust for all image sizes including low-resolution images. Existing high-to-low and low-to-high frameworks [6-9, 18-19] use a single resolution input image. In these methods, the bounding box images of various sizes for target human are scaled to a large size such as 256×192 or 384×288 , and then insert them into the network. In their methods, the size of the input images is fixed for all target images with different resolutions. Therefore, for the low-resolution target images, artifacts occur due to up-sampling. The degraded image adversely affects accurate estimation of the joint positions. To resolve this problem, in the proposed method, the bounding box image for target human is rescaled to multiple input images with various sizes, and the features extracted from the multiple images are fused in the network. In this way, the network can refer both high-resolution and low-resolution input features, which leads to an accurate pose estimation.

The bounding box image for target person I , whose height and width are h and w , is rescaled to multi-resolution input images $I_1, I_2, \dots, I_n, \dots, I_N$, where the sizes of the rescaled images are $(h_1 \times w_1)$, $(h_2 \times w_2)$, \dots , $(h_N \times w_N)$ respectively. They are generated by bicubic interpolation of I . As illustrated in Fig. 2, the stem network consists of two strided convolution layers that extract features from the input image I_n . The extracted feature of I_n is inserted into S_{n1} . For example, the extracted feature map of I_1 becomes the input of S_{11} . The feature from the middle layer of the upper subnetwork is fused with the feature extracted from the stem network. To match the dimension of the features, the feature map from the upper subnetwork is downsampled. Assuming that the feature map from the $(n-1)$ -th subnetwork are fused to the input of n -th subnetwork, the input feature of S_{n1} can be expressed as follows:

$$\text{Input feature of } S_{n1} = \text{stem}(I_n) \oplus \text{conv}(S_{(n-1)1}(\text{stem}(I_{n-1})), \text{stride} = 2)$$

3.2.2 Guiding Channel

We introduce a guiding channel, which induces the multi-resolution input features to

alternatively affect the network according to the resolution of the original target image. When the target image is low-resolution, the network is induced to be influenced by the image scaled up at a small magnification which has little degradation. On the other hand, when the target image is high-resolution, it is induced to be affected by the large-sized input image which does not lose much information.

As illustrated in Fig. 2, for the 3-channel input image I_n , a $h_n \times w_n$ matrix whose elements are binary values is added as the 4th channel of I_n , which is called a guiding channel. The guiding channel is applied in $I_n, n > 1$. With the guiding channel added, the input of the stem network becomes a 4-channel image. When the resolution of the target image is within the predefined range, all element values of the guiding channel are filled with 1, whereas when the resolution is out of the range, it is filled with 0. The performance was additionally increased by the introduction of the guiding channel, and related experiments are described in section 4.3.

3.2.3 Variant Architectures

The experiment was conducted with three model variations as shown in Table 1 to verify the effect of the number of input images and their rescaling resolutions. HRNet [9] includes HRNet-W32, a model with fewer parameters, and HRNet-W48, a model with more parameters. Experiments were conducted using HRNet-W32 as a baseline network. The resolution of I_1 is set to 256×192 , which is the resolution of the input image in HRNet-W32. Additionally, I_2 or I_3 scaled to the lower resolution is inserted into the other subnetworks. The last column of Table 1 indicates the conditions under which each element of the guiding channel has a values of 1 or 0. Fig. 3 shows the structures of the model variations corresponding to (a), (b), and (c) shown in Table 1. The performances of all three models surpass the performance of the baseline on the low-resolution test images. A detailed explanation of the results is described in Section 4.3.

Table 1. Proposed three model variations with multi-resolution input images.

	I_n (height x width)	Input feature (height x width)	Guiding Channel (condition)
OURS-(a)	I_1 (256×192)	64×48	-
	I_2 (64×48)	32×24	1, if $96 \cdot 96 < h \cdot w$ 0, otherwise
OURS-(b)	I_1 (256×192)	64×48	-
	I_3 (32×24)	16×12	1, if $96 \cdot 96 < h \cdot w$ 0, otherwise
OURS-(c)	I_1 (256×192)	64×48	-
	I_2 (128×96)	32×24	1, if $64 \cdot 64 < h \cdot w < 128 \cdot 128$ 0, otherwise
	I_3 (64×48)	16×12	1, if $h \cdot w < 64 \cdot 64$ 0, otherwise

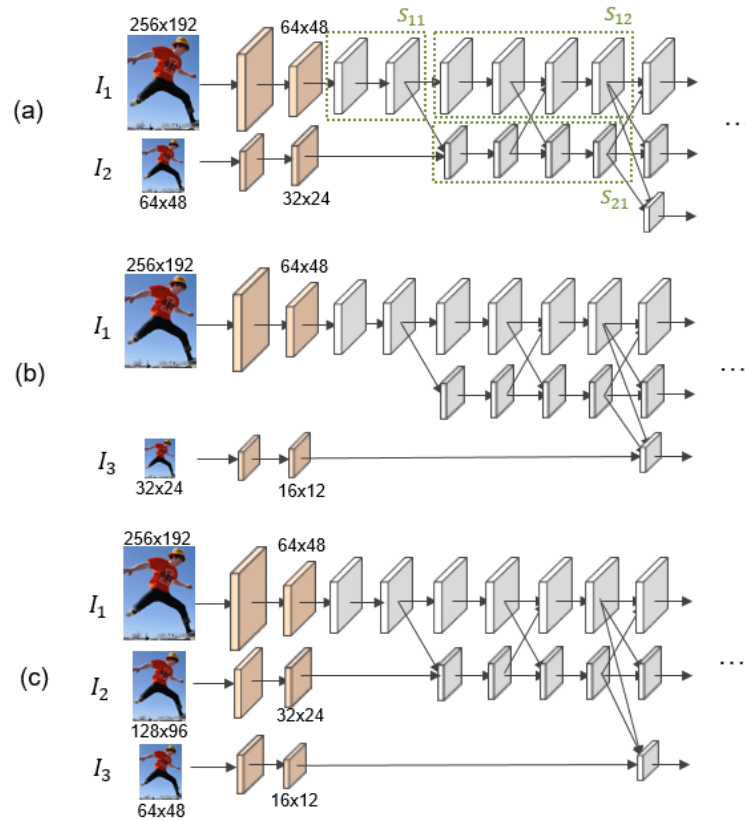


Fig. 3. Structures of the model variations corresponding to (a), (b), and (c) in [Table 1](#).

4. Experimental Result

4.1 Dataset

There is no publicly available pose estimation dataset for a distant person or low-resolution images. Instead, the most famous pose estimation dataset, MS COCO [21] is used for training and evaluation. In the COCO [21] dataset, the skeleton of a person is defined as 17 joint keypoints. It contains 57K images and more than 150K human instances. As shown in [Fig. 4](#), the distribution of the bounding boxes in the COCO dataset is biased toward high resolution. To obtain a sufficient number of low-resolution images for an evaluation set, we extracted samples with a bounding box area larger than 64×64 and downsampled them to 24×18 , 32×24 , 48×36 , 64×48 size images.

4.2 Evaluation Protocols

For a fair comparison with the baseline, we follow the data augmentation of HRNet [9]. To operate robustly on images of varying resolutions, the network is learned from the images scaled from 0.65 to 1.35 times the original size of the bounding box. In addition, random rotation ($[-45^\circ, 45^\circ]$) and random flip are applied to the training images. The network is trained by Adam optimizer with the base learning rate $1e-3$. The learning rate drops to $1e-4$ and $1e-5$ at 170 and 200 epochs, respectively. There are 210 epochs in total.

For the evaluation, the similarity between the estimated joint position and the corresponding

ground truth is calculated by Object Keypoint Similarity (OKS). OKS is represented as

$$\text{OKS} = \frac{\sum_i \exp(-d_i^2 / 2s^2k_i^2) \delta(v_i > 0)}{\sum_i \delta(v_i > 0)}$$

where d_i denotes the Euclidean distance between two joints and v_i the visibility flag on the ground truth. s and k_i denotes the human size and a per-keypoint constant that controls falloff, respectively. If the OKS value is greater than the predefined threshold, the joint estimation is considered a success. The performance is measured by average precision (AP) and average recall (AR) of all joints. AP at threshold 0.5 and 0.75 are expressed as $AP^{0.5}$ and $AP^{0.75}$, respectively. And we use an average value of AP at thresholds [0.5, 0.55, ..., 0.95] as one of the evaluation metrics and denote it as AP . Similarly, we calculate AR which is the average value of AR at thresholds [0.5, 0.55, ..., 0.95].

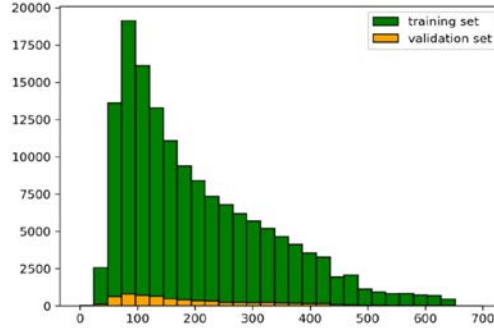


Fig. 4. Human bounding box resolution distributions in the COCO training and validation datasets. The horizontal and vertical axes denote the image height and the number of bounding boxes belonging to the bin, respectively. The COCO dataset consists of high-resolution images, and there are few images with a height of 100 pixels or less.

4.3 Ablation study

An ablation study was conducted to investigate the effect of the multi-resolution input structure and the guiding channel. **Table 2** shows the pose estimation results of our proposed method without the guiding channel, with the guiding channel, and the baseline. The method without the guiding channel leverages the multi-resolution inputs which consist of RGB three channels. The results are obtained for 5000 human bounding boxes in the COCO validation set. **Table 2** shows the AP of each target resolution. As a result of using the multi-resolution structure, the overall performance exceeds the baseline. In particular, the results at 24×18 and 32×24 showed significant improvements, about 1.5 AP scores, compared to the baseline. In the case of using the guiding channel, the performance improved about 0.4 to 0.6 AP scores in all scales. It is verified that leveraging a less degraded image in addition to the 256×192 size input image used in the HRNet [9] helps to accurately estimate poses from low-resolution images. In addition, the guiding channel that provides information about which of the input images is closest to the target image size contributes to performance improvement as well.

Next, we conducted experiments with our proposed model variations explained in section 3.2.3 to find the optimal model structure. **Table 3** shows the experimental results for model variations (a), (b), and (c) in **Table 1** and **Fig. 3**. All three models outperform the baseline.

Model (a) shows the best performance at 48×36 and 64×48, on the other hand, model (c) shows the best performance at 24×18, 32×24, and 48×36 which proves the usefulness of using multi-resolution input features.

4.4 Comparison on MSCOCO Dataset

Table 2. Comparison with baseline and our proposed methods. Results are *APs* (average AP at thresholds [0.5, 0.55, ..., 0.95]) on COCO validation set.

Test bounding box size (height × width)	24×18	32×24	48×36	64×48
Baseline	18.9	37.1	59.0	68.8
OURS (without guiding channel)	20.4	38.6	59.7	69.0
OURS (with guiding channel)	21.0	39.0	60.2	69.5

Table 3. Comparison with baseline and our model variations explained in section 3.2.3. Results are *APs* (average AP at thresholds [0.5, 0.55, ..., 0.95]) on COCO validation set.

Test bounding box size (height × width)	24×18	32×24	48×36	64×48
Baseline	18.9	37.1	59.0	68.8
OURS-(a)	21.0	39.0	60.2	69.5
OURS-(b)	21.0	39.0	59.9	69.2
OURS-(c)	21.3	39.1	60.2	69.1

Table 4. Comparison with other 2D human pose estimation baselines on COCO validation set.

	backbone	<i>AP</i>	<i>AP</i> ⁵⁰	<i>AP</i> ⁷⁵	<i>AP</i> ^M	<i>AP</i> ^L	<i>AR</i>
Hourglass [6]	8-stage Hourglass	66.9	-	-	-	-	-
CPN [8]	ResNet-50	68.6	-	-	-	-	-
SimpleBaseline [7]	ResNet-50	70.4	88.6	78.3	67.1	77.2	76.3
SimpleBaseline [7]	ResNet-152	72.0	89.3	79.8	68.7	78.9	77.8
HRNet [9]	HRNet-W32	74.4	90.5	81.9	70.8	81.0	79.8
OURS-(a)	HRNet-W32	74.3	90.2	82.1	70.7	81.2	79.7

We tested the proposed network using the entire COCO validate image set including both small and large images. The performance of the proposed network is competitive with the baseline as shown in Table 4. Both *AP* and *AR* showed the competitive performance to the baseline, HRNet [9]. The input image size of all the compared baseline networks is 256×192. The experimental result verifies that the proposed model is not only targeted for low-resolution images but rather improves accuracy on low-resolution images while maintaining the performance on high-resolution images. Note that, because the high-resolution image dominates the distribution of the COCO validation dataset, the performance improvement of the low-resolution images did not significantly affect the overall average score.

Additionally, we investigate the effect of using multi-resolution input images for a backbone network with large parameters, which is the HRNet big model with an input size of 384×288 [9]. We represent the proposed model structure in Table 5. Table 6 shows the comparison with baseline and our proposed method. HRNet, with the input size 384×288 yields severe performance degradation on low-resolution images than HRNet with input size 256×192. When the proposed method is applied to the big HRNet network, the performance on low-

resolution images significantly increases. In particular, in 24×18 and 32×24 , the performance increases by 4.1 and 2.9 AP scores, respectively, which is a larger gap than the case of HRNet 256×192 backbone.

The qualitative evaluation results of the baseline and the proposed methods are illustrated in **Table 6-8**. First, the results on the test set made by scaling the images of the COCO validation set to 24×18 size is shown in **Table 7**. The proposed model estimates the joint positions more accurately compared to the baseline, and the skeleton represents a pose similar to the ground truth. **Table 8** shows the results of the test images with 32×24 , 48×36 , and 64×48 resolutions. **Table 9** illustrates the comparison of the estimated pose of the baseline and the proposed method as the human bounding box resolution increases, for the difficult image sample. In 24×18 size, the upper body and the background are hard to distinguish even with the eye, and both methods estimate the background parts as joints, resulting in a large error. In 32×24 size, our method accurately estimates the joints except for both arms. On the other hand, a large error still occurs in the baseline method. For the 64×48 size, our method yields more accurate results on in the right arm compared to the baseline.

The qualitative evaluation results of the baseline and the proposed methods are illustrated in **Table 7-9**. First, the results on the test set made by downsampling the images of the COCO validation set to 24×18 size are shown in **Table 6**. The proposed model estimated the joint position relatively accurately compared to the baseline, and the skeleton represents a pose similar to the ground truth. **Table 7** shows the estimated poses for the test images with 32×24 , 48×36 , and 64×48 resolutions. **Table 8** illustrates the comparison of the estimated poses of the baseline and that of the proposed method as the human bounding box resolution increases, for the difficult image sample. In 24×18 size, the upper body and the background are hard to distinguish, therefore, both methods result in a large error. In 32×24 size, our method accurately estimates the joints except for both arms. On the other hand, a large error still occurs in the baseline method. For the 64×48 size, our method yields more accurate results on the right arm compared to the baseline.


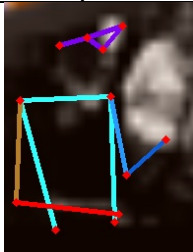
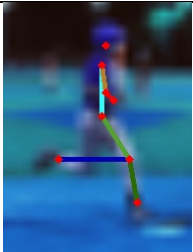
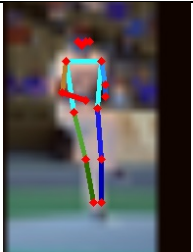

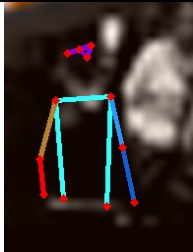
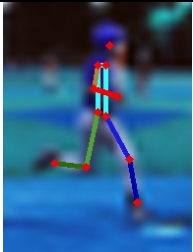
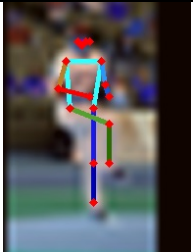
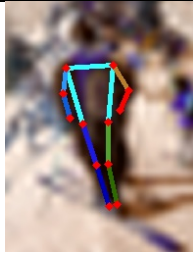
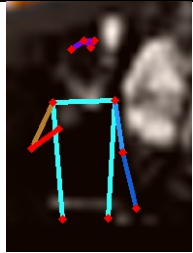
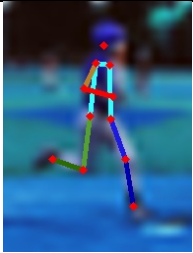
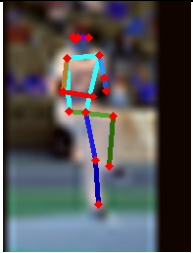
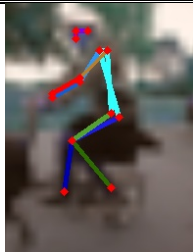
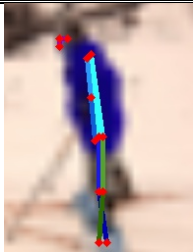
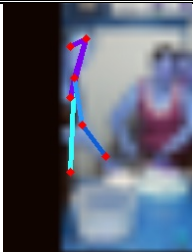
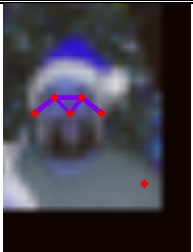


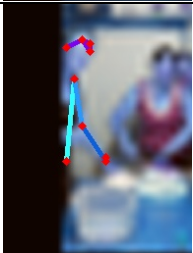
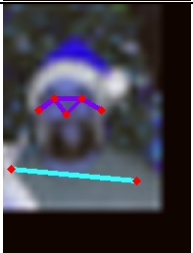
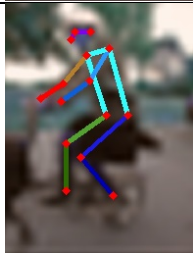
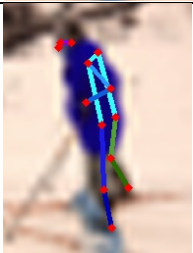
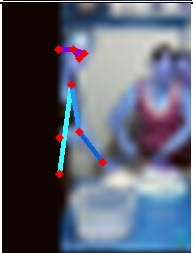
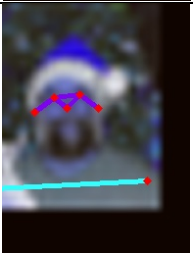
Table 5. Proposed model structure using the backbone of HRNet-W48.

I _n (height × width)	Input feature (height × width)	Guiding Channel (condition)
I1 (384×288)	96×72	-
I2 (96×72)	48×36	1, if $96 \cdot 96 < w \cdot h$ 0, otherwise

Table 6. Comparison with baseline and our proposed method using the backbone of HRNet-W48. Results are APs (average AP at thresholds [0.5, 0.55, ..., 0.95]) on COCO validation set.

Test bounding box size (height x width)	24×18	32×24	48×36	64×48
Baseline	15.8	35.6	59.1	69.5
OURS	19.9	38.5	60.8	70.2

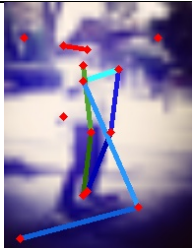
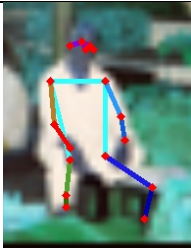


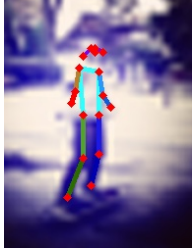
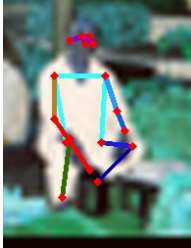

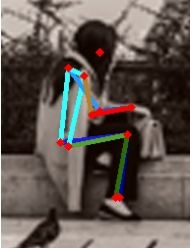
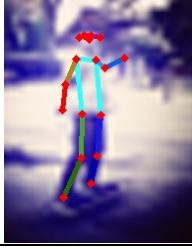
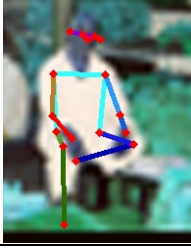


Table 7. Comparison of pose estimation results with the baseline on the COCO validation set downsampled to 24×18 size.

Baseline				
OURS				
GT				
Baseline				
OURS				
GT				

4.5 Applications

We performed additional experiments by applying the proposed method to the images captured from an actual surveillance camera system. The test videos were recorded with an SD camera on top of a building. It is difficult to estimate the human pose in frames extracted from the video because they are low resolution and the shooting angle is slanted. The model weights learned from the COCO training images were used. As illustrated in Fig. 5, the baseline incorrectly detects body joints from the shadow part, on the other hand, the proposed method yields relatively accurate results

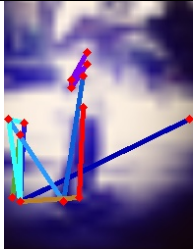
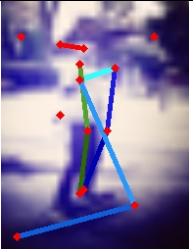
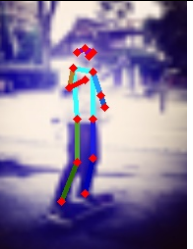

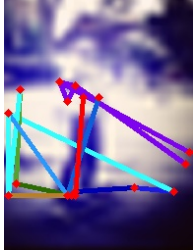
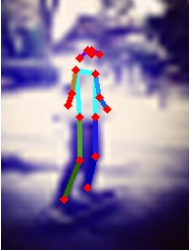
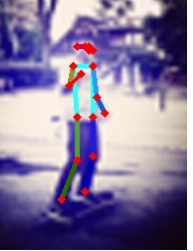

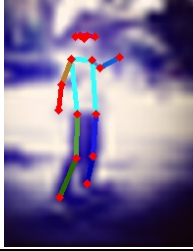
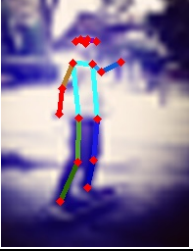


Table 8. Comparison of pose estimation results with the baseline on the COCO validation set downsampled to 32×24 , 48×36 , and 64×48 sizes.

	32×24	32×24	48×36	64×48
Baseline				
OURS				
GT				

5. Conclusions

In this paper, we have proposed a multi-resolution input feature fusion network that improves accuracy in low-resolution images while maintaining the performance for images of various sizes. In the proposed method, the bounding box human image is scaled to multiple input images of various sizes and inserted into the network. The stem network extract features from the input images and the features are fused between the subnetworks. Moreover, we have devised a guiding channel which induces the multiple input rescaled images to alternatively affect the network according to the resolution of the original target image. The experiments on MS COCO dataset and videos obtained from the actual surveillance camera system have demonstrated the effectiveness of our approach. Our method outperforms previous studies in

Table 9. Comparison of the estimated poses of the baseline and that of the proposed method as the human bonding box resolution increases, for the difficult image sample. Tested resolutions are 24×18, 32×24, 48×36, and 64×48.

	24×18	32×24	48×36	64×48
Baseline				
OURS				
GT				

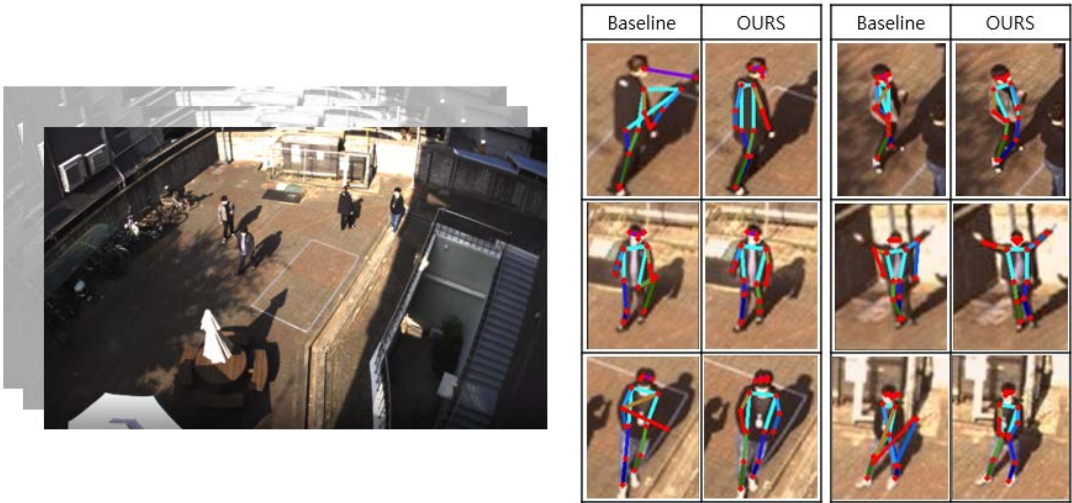


Fig. 5. Test images captured from an actual surveillance camera system (left). The test videos were recorded with an SD camera on top of a building. Qualitative results of our proposed method and the baseline (right). The proposed method yields relatively accurate results to the baseline.

low-resolution images, however, estimation errors still exist when the boundary between the person and the background is blurred or the colors of the person and the background are similar. In this manner, research to estimate accurate joint positions of an image in which the distinction between a person and a background is ambiguous will be a meaningful future study.

Acknowledgement

This work was partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 1711159681, Development of high-quality AI-AR interactive media service through deep learning-based human model generation technology) and Basic Research Program of Korea Electronics Technology Institute (KETI)).

References

- [1] N. Dalal, and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. of 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognitions (CVPR)*, pp. 886-893, 2005. [Article \(CrossRef Link\)](#).
- [2] D. C. He, and L. Wang, "Texture Unit, Texture Spectrum, And Texture Analysis," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 28, no. 4, pp. 509-512, July 1990. [Article \(CrossRef Link\)](#).
- [3] X. Wang, T. X. Han, S. Yan, "An HOG-LBP human detector with partial occlusion handling," in *Proc. of International Conference on Computer Vision (ICCV)*, pp. 32-39, 2009. [Article \(CrossRef Link\)](#).
- [4] Z. Cao, T. Simon, S. Wei, and Y. Sheikh, "Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields," in *Proc. of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1302-1310, 2017. [Article \(CrossRef Link\)](#).
- [5] M. Kocabas, S. Karagoz, and E. Akbas, "MultiPoseNet: Fast Multi-Person Pose Estimation using Pose Residual Network," in *Proc. of European Conference on Computer Vision (ECCV)*, pp. 437-453, 2018. [Article \(CrossRef Link\)](#).
- [6] A. Newell, K. Yang, and J. Deng, "Stacked Hourglass Networks for Human Pose Estimation," in *Proc. of European Conference on Computer Vision (ECCV)*, pp. 483-499, 2016. [Article \(CrossRef Link\)](#).
- [7] X. Bin, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," in *Proc. of the European conference on computer vision (ECCV)*, 2018.
- [8] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun, "Cascade Pyramid Network for Multi-Person Pose Estimation," in *Proc. of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7103-7112, 2018. [Article \(CrossRef Link\)](#).
- [9] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep High-Resolution Representation Learning for Human Pose Estimation," in *Proc. of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5686-5696, 2019. [Article \(CrossRef Link\)](#).
- [10] J. Huang, Z. Zhu, F. Guo, and G. Huang, "The devil is in the details: Delving into unbiased data processing for human pose estimation," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5700-5709, 2020. [Article \(CrossRef Link\)](#).
- [11] Y. Bai, Y. Zhang, M. Ding, and B. Ghanem, "Finding Tiny Faces in the wild with Generative Adversarial Network," in *Proc. of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21-30, 2018. [Article \(CrossRef Link\)](#).
- [12] Y. Bai, Y. Zhang, M. Ding, and B. Ghanem, "SOD-MTGAN: Small Object Detection via Multi-Task Generative Adversarial Network," in *Proc. of European Conference on Computer Vision (ECCV)*, pp. 210-226, 2018. [Article \(CrossRef Link\)](#).

- [13] L. Neumann, A. Vedaldi, "Tiny people pose," in *Proc. of Asian Conference on Computer Vision (ACCV)*, vol. 11363, pp. 558-574, May. 2018. [Article \(CrossRef Link\)](#).
- [14] X. Xu, H. Chen, F. Moreno-Noguer, L. A. Jeni, and F. De la Torre, "3D Human Shape and Pose from a Single Low-Resolution Image with Self-Supervised Learning," in *Proc. of European Conference on Computer Vision (ECCV)*, pp. 284-300, 2020. [Article \(CrossRef Link\)](#).
- [15] Z. Zhang, L. Wan, W. Xu, and S. Wang, "Estimating a 2D pose from a tiny person image with super-resolution reconstruction," *Elsevier Computers & Electrical Engineering*, vol. 93, July, 2021. [Article \(CrossRef Link\)](#).
- [16] M. Li, Z. Zhou, J. Li, and X. Liu, "Bottom-up Pose Estimation of Multiple Person with Bounding Box Constraint," in *Proc. of International Conference on Pattern Recognition (ICPR)*, pp. 115-120, 2018. [Article \(CrossRef Link\)](#).
- [17] A. Newell, A. Huang, and J. Deng, "Associative Embedding: End-to-End Learning for Joint Detection and Grouping," *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 2274-2284, 2017.
- [18] S. Park, M. Ji, and J. Chun, "2D human pose estimation based on object detection using RGB-D information," *KSII Transactions on Internet and Information Systems (TIIS)*, 12(2), 800-816, 2018. [Article \(CrossRef Link\)](#).
- [19] N. Yali, J. Lee, S. Yoon, and D. S. Park, "A Multi-Stage Convolution Machine with Scaling and Dilation for Human Pose Estimation," *KSII Transactions on Internet and Information Systems (TIIS)*, 13(6), 3182-3198, 2019. [Article \(CrossRef Link\)](#).
- [20] S. Liu, G. Hua, and Y. Li, "2.5D human pose estimation for shadow puppet animation," *KSII Transactions on Internet and Information Systems (TIIS)*, 13(4), 2042-2059, 2019. [Article \(CrossRef Link\)](#).
- [21] 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. of European conference on computer vision (ECCV)*, pp. 740-755, 2014. [Article \(CrossRef Link\)](#).
- [22] A. Mykhaylo, "2d human pose estimation: New benchmark and state of the art analysis," in *Proc. of the IEEE Conference on computer Vision and Pattern Recognition (CVPR)*, 2014. [Article \(CrossRef Link\)](#).
- [23] Li, Zhigang, et al, "Temporal and Spatial Traffic Analysis Based on Human Mobility for Energy Efficient Cellular Network," *KSII Transactions on Internet and Information Systems (TIIS)*, 15(1), 114-130, 2021. [Article \(CrossRef Link\)](#).
- [24] Ma, Ruoxin, Shengjie Zhao, and Samuel Cheng, "Self-Supervised Rigid Registration for Small Images," *KSII Transactions on Internet and Information Systems (TIIS)*, 15(1), 180-194, 2021. [Article \(CrossRef Link\)](#).
- [25] Zhao, Liquan, and Yupeng Zhang, "Generative Adversarial Networks for single image with high quality image," *KSII TRANSACTIONS ON INTERNET AND INFORMATION SYSTEMS*, 15(12), 4326-4344, 2021. [Article \(CrossRef Link\)](#).
- [26] Dong, Xiang, et al, "Dual Attention Based Image Pyramid Network for Object Detection," *KSII TRANSACTIONS ON INTERNET AND INFORMATION SYSTEMS*, 15(12), 4439-4455, 2021. [Article \(CrossRef Link\)](#).



Boeun Kim received the B.S degree in Electronic and Electrical Engineering from Korea Advanced Institute of Science and Technology. She received the M.S degree and is currently pursuing a Ph.D. degree in Electrical and Computer Engineering at Seoul National University. From 2015 to 2017, she was a research engineer in Samsung Electronics. Since 2017, she has been with Korea Electronics Technology Institute. Her research interests include deep learning and human motion pattern analysis.



Yeonseung Choo was born in Suwon, Korea, in 1993. He received his BSc in Information and communication Engineering from Sunmoon University, Korea, in 2018. And he received an MSc in Image Science at Chung-Ang University. His research interests include video stitching and video registration.



Hea In Jeong received B.S. and M.S. degrees in Department of Computer Science from Sookmyung Women's University, Korea. She is currently an assistant researcher at Korea Electronics Technology Institute. Her research interests include 3D Displays, Image Processing, Computer Vision, and Machine Learning.



Chung-II Kim received his B.S. and M.S. degrees in Electronic Engineering from Korea University, Seoul, Korea, in 2018 and 2020, respectively. He is currently a researcher at the Artificial Intelligence Research Center, Korea Electronics Technology Institute, Gyeonggi-do, Korea. His current research interests include image processing, signal processing, and generative model.



Saim Shin received her B.E. degree in Computer Science from Sookmyung Women's University in 2000, M.E. degree in Computer Science from KAIST (Korea Advanced Institute of Science and Technology) in 2002, and Ph.D. degree from the University of Sogang in 2018. Since 2006, she has been working at Korea Electronics Technology Institute (KETI), and she is presently in charge of the Artificial intelligence research center as a principal researcher and a director. Her research interests include vision and natural language processing, and machine learning.



Jungho Kim received his B.S. degree from the Kyungpook National University, Korea, in 2004 and M.S. and Ph.D. degrees from Korea Advanced Institute of Science and Technology (KAIST) in 2006 and 2013, respectively. He is now a principal researcher at Korea Electronics Technology Institute (KETI). He has general research interests in computer vision, vision-based robotics and statistical inference. He has published papers on SLAM, object tracking and vision-based autonomous navigation.