Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 3447827, 17 pages
https://doi.org/10.1155/2022/3447827

Research Article

A Structure-Aware Adversarial Framework with the Keypoint Biorientation Field for Multiperson Pose Estimation

Xianjia Meng,[1] **Yong Yang**,[1] **Kang Li**,[1] **and Zuobin Ying**[2]

[1]*School of Information Science and Technology, Northwest University, Xi'an 710119, China*
[2]*Institute of Data Science, City University of Macau, Macao, Macau*

Correspondence should be addressed to Kang Li; likang@nwu.edu.cn

Received 12 August 2021; Accepted 13 November 2021; Published 14 February 2022

Academic Editor: Ximeng Liu

Human pose estimation is aimed at locating the anatomical parts or keypoints of the human body and is regarded as a core component in obtaining detailed human understanding in images or videos. However, the occlusion and overlap upon human bodies and complex backgrounds often result in implausible pose predictions. To address the problem, we propose a structure-aware adversarial framework, which combines cues of local joint interconnectivity and priors about the holistic structure of human bodies, achieving high-quality results for multiperson human pose estimation. Effective learning of such cues and priors is typically a challenge. The presented framework uses a nonparametric representation, which is referred to as the Keypoint Biorientation Field (KBOF), to learn orientation cues of joint interinteractivity in the image, just as human vision can explore geometric constraints of joint interconnectivity. Additionally, a module using multiscale feature representation with inflated convolution for joint heatmap detection and Keypoint Biorientation Field detection is applied in our framework to fully explore the local features of joint points and the bidirectional connectivity between them at the microscopic level. Finally, we employ improving generative adversarial networks which use KBOF and multiscale feature extraction that implicitly leverages the cues and priors about the structure of human bodies for global structural inference. The adversarial network enables our framework to combine information about the connections between local body joints at the microscopic level and the structural priors of the human body at the global level, thus enhancing the performance of our framework. The effectiveness and robustness of the network are evaluated on the task of human pose prediction in two widely used benchmark datasets, i.e., MPII and COCO datasets. Our approach outperforms the state-of-the-art methods, especially in the case of complex scenes. Our method achieves an improvement of 2.6% and 1.7% compared to the latest method on the MPII test set and COCO validation set, respectively.

1. Introduction

Human body pose estimation is widely researched in the field of computer vision, which involves the positioning and pose configuration of human body parts and the multimedia data captured by sensors, especially images and videos. Human body geometry information and movement information provided by human body pose estimation are widely used in many fields, e.g., behavior recognition, motion prediction, human-computer interaction, virtual reality, and rehabilitation training. At the same time, there are now some specific application scenarios. For example, smart manufacturing factories can recognize human actions and intentions based on the industrial Internet through human body gesture recognition and prediction [1]. Home care robots can detect accidents by detecting human body posture [2].

In recent years, tremendous progress has been achieved in this field due to the powerful feature extraction and reasoning capabilities of Deep Convolutional Neural Networks (DCNNs) [3–5]. These methods mainly predict the heatmaps of the joint points of the human body and show good feature extraction and representation capabilities. However, the main challenge of reasoning about the human pose, especially in multiperson scenes, is the flexibility of the human pose and physical occlusion caused by foreign

objects and the weak robustness to partial joint recognition caused by complex scenes. These models may generate implausible heatmaps of predicted joints when faced with the above challenging scenarios. In these scenarios, the network may be forced to learn features similar to human joints. These features may be in the background image or belong to another person, as shown in Figure 1(a).

One of the direct and effective ways to address these challenging scenarios is to incorporate the priors of human structure into the training process of the network to make the pose predictions more reasonable. Generative adversarial networks (GANs) have been applied to network training to allow the network to learn the structural constraints between body parts [6, 7]. However, this has some shortcomings for pose prediction models based only on traditional GANs, because the combination of the priors of human structure into the network is based on adequate joint detection and connection relations between joints based on local features. If joint detection and connectivity relationships between joints at the microlevel are not taken into account, previous GAN-based approaches [6, 7] still cannot adequately address these challenging scenarios when more complex body part occlusions and more intrusive backgrounds occur.

We can reasonably infer the position of the joint points from human vision based on the observed connections between the local joint points combined with the priors of body structure. Even under severe occlusion and interference, we can infer plausible poses. However, how to make the network learn to reason is a challenge for the interconnections between local articulations and oriented information to help the network to infer from one articulation to another. Inspired by [8, 9] and combined with human vision's reasoning process, we propose the Keypoint Biorientation Field (KBOF), as shown in Figures 1(c) and 1(d). It can help the network learn the bidirectional and location information between the nodes based on the local features of the pictures at the microlevel and guide the network to generate a more accurate heatmap of the human nodes. In addition, to fully detect the local features of the articulation points and the bidirectional relationship between them at the microlevel, we apply the multiscale feature representation widely used in semantic segmentation tasks [10, 11] to the network. Combining the above two parts, a module using multiscale feature representation for joint heatmap and KBOF detection is applied to our framework. Finally, to embed the priors of human structure into the network, a discriminator is proposed to dominantly check whether the human structure prior is plausible or not. This discriminator follows the same network structure as the generator [12]. The discriminator can distinguish plausible pose configurations from implausible ones and thus guide the generation to iterate towards generating more accurate ones. After training is completed, the discriminator is removed and the generator is used as a pose predictor.

The three main contributions of this work are as follows. First, we design an adversarial network for multiperson pose recognition, which takes the human structural geometric prior into account. By embedding the human structure prior in the network, the predicted joint heatmap errors can be effectively reduced for occlusion and foreign object interference in complex environments. Second, we present the Keypoint Biorientation Field (KBOF), which can encode bidirectional and positional information between local articulation points to guide the network to generate more reliable articulation heatmaps. Third, we present a multiscale dilated convolution module to increase the field of view of the network. The module can improve the performance of the network by exploring multiscale local features at the microlevel for the network.

2. Related Work

Our approach is closely related to work based on CNNs using heatmaps for the human pose, multiscale feature representation networks, and generative adversarial networks.

2.1. Human Pose Estimation. Traditional 2D HPE approaches use probabilistic graphical models [13, 14] and pictorial structural models with hand-drawn image features [15, 16] to detect body parts. Although improvements can be made by these clever model designs and algorithmic implementations [17], the bottleneck seems to be the lack of effective feature representations that can characterize visual cues at different scales and vary according to different character profiles and environments. The popularity of DCNNs has changed this situation in computer vision. Significant improvements were realized in the field [3, 8, 18–20] in earlier research works.

The convolutional pose machine [19], one of the most popular deep learning-based methods, presents a sequential framework that increasingly refined keypoint heatmap estimation through a series of stages throughout the network. Built upon [19], Cao et al. combined the idea of sequential refinement with the newly proposed Part Affinity Field (PAF) result in the OpenPose method [8]. Further, based on PAF, Cao et al. [7] proposed the composite field, which consists of Part Intensity Field (PIF) and Part Association Field (PAF) whose functions are to locate the nodes and connect the localized nodes, respectively. Our multiscale feature extraction module draws on the ideas in [8, 9] and proposes a bidirectional representation of the orientation and location information between two keypoints to enhance the reasoning.

Multiscale feature extraction has been widely applied in DCNNs for pose recognition. Newell et al. proposed stacked hourglass networks [21] pooling and upsampling in successive steps to capture the diverse space of associations between body parts to generate the final prediction. Based on the hourglass structure, Chu et al. embedded a multicontext approach with an attention mechanism into the hourglass backbone [22] for human pose recognition. The hourglass residual units are superimposed on an origin network incorporating the holistic attention model, which has the advantage of a multiscale field of view. Conditional random fields are used in the postprocessing stage to connect the detected nodes. However, this method has the disadvantage of increasing the computational cost and significantly increasing the inference time.

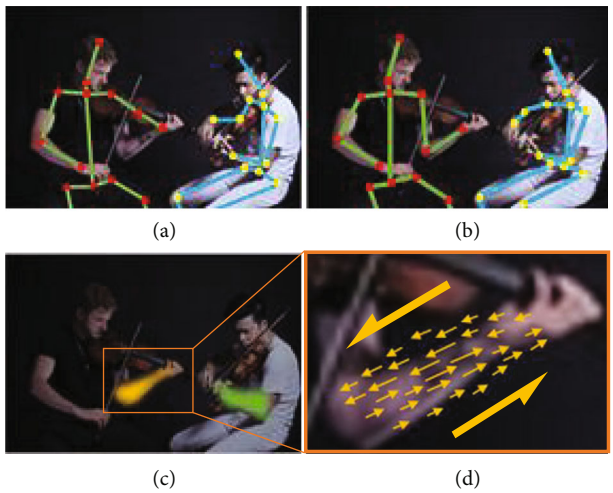


Figure 1: (a) Incorrect postures may be generated by ordinary adversarial networks without contact information between joint points. (b) Combining the connection information between local joints and the prior body structure results in more accurate predictions. (c) Keypoint Biorientation Field (KBOF) corresponds to the contact information between two joints. (d) A pair of bidirectional vectors in each pixel of every KBOF encodes the orientation and position information between two keypoints.

To address the detection of "hard" keypoints in challenging scenarios, Chen et al. proposed the Cascaded Pyramid Network (CPN). CPN first uses GlobalNet to locate common "simple" keypoints followed by a RefineNet to explicitly identify "hard" keypoints by integrating all levels of feature representation and combining keypoint missing losses. Enhanced CPN [23] was proposed with a shuffle unit to fuse feature maps from the pyramid and an attention module to extract more representative feature maps for pose tracking.

Unlike many approaches that recover the high-resolution representation through a low-to-high decoding network, the high-resolution network [20] maintains the high-resolution feature representation across all network phases. It starts with a high-resolution subnetwork in the first stage. High-to-low resolution subnetworks are added sequentially, and multiresolution subnetworks are connected in a parallel manner. This approach of maintaining high resolution while increasing the resolution at different levels can accept more semantic information leading to rich feature representations. HigherHRNet [24] was proposed to address the challenge of correctly identifying the size of small humans in scale variation. Its combination of feature maps output from HRNet and high-resolution representations from upsampling by transposition convolution can address the scale variation challenge. A similar approach to HRNet was used in [25], which combines cross-stage feature aggregation and progressively refined intermediate supervision to achieve performance gains at a constant computational cost. Our approach combines the enhanced HRNet of [20, 24] with a dilated convolution in the backbone network with multiple dilated rates to extract multiscale features for locating keypoints and exploring the bidirectional information between them.

2.2. Multiscale Feature Extraction. The concept of multiscale feature representation can be traced back to the theory of scale space [26]. In recent years, multiscale feature extraction based on CNNs has been widely used in several task areas in computer vision. Most approaches usually use feature pyramids to perform multiscale feature extraction for target detection tasks. Feature pyramid networks are representative works that construct feature pyramids by multiscale feature aggregation. Kim et al. proposed parallel feature pyramid networks (CPNs) [27], which can overcome the limitations of the original multifeature extraction; i.e., the different abstraction layers of feature layers of CNNs limit the detection performance. The method uses spatial pyramid pooling and feature transformation to generate different-sized feature maps through widening the network. The feature maps at each scale are subsequently rescaled to a uniform size and combined with global information to generate the final feature pyramid.

The design flaws of CPN, which limit the full utilization of multiscale features, were identified by Guo et al. who proposed AugFPN [10] to address these problems. AugFPN has three different modules: consistent supervision, residual feature enhancement, and ROI selection, corresponding to the three flaws of FPN. The prominent advancement is consistent supervision, which narrows the semantic divide between features at different scales before aggregation. In Big-Little Net [28], multiscale branching networks are used, which have different computational efficiencies with different resolutions in different branches. By fusing features from different branches at different scales, the model can obtain multiscale features with less inference time.

Another multiscale feature extraction method, which refers to dilated convolution or atrous convolution, is fundamentally different from the above methods. Dilated convolution [29] can explicitly adjust the field of view of the subnetwork module by controlling the resolution of the feature map with different dilated ratios to extract features from different scales of field of view. It increases the resolution without increasing the number of parameters and keeping the computational efficiency constant. Dilated convolution is used to capture multiscale contextual features with multiple dilation rates arranged in cascade or parallel. In DeepLab [30], dilated convolution is used for upsampling filters to effectively expand the field of view of the filters while keeping the network size constant and computational efficiency unaffected. Dilated convolution is often used for feature extraction at the microscopic level. In DetectoRS [11], switchable dilation convolution is used to convolve different features with different dilation ratios and use switch functions to aggregate the results. This significantly improves the performance of the method for target detection tasks.

Taking all these considerations into account, dilated convolution is used in our framework since KBOF explores the bidirectional detail information between keypoints, which has high requirements for feature extraction at the microlevel relative to the body structure. The multiscale feature representation is improved on keypoint heatmaps and KBOF in [30], which combines a parallel architecture where multiscale features are first processed through a filter and

branch out into multiple parallel streams with different dilated convolution rates. It goes beyond the common cascade approach by combining multiple parallel streams and averaging the original input pooling to achieve detailed feature representation at the microscopic level.

2.3. Generative Adversarial Networks. Generative adversarial networks (GANs) have been widely used in several scenarios, such as generating sample images from i.i.d. data given unknown distributions [31], visualization operations generated in natural manifold [32], generating images of people from given pose [33], and data augmentation for human pose recognition tasks [34]. Since the first GAN for learning the possible distribution of training samples and generating similar products [35] was proposed, research on GAN has focused on three prospects: (1) generation of high-quality images, (2) stable training, and (3) exploration of scenarios for GAN use.

Deep Convolutional GAN [36] first applied the deconvolution operation to generators. DCGAN has several key modifications that are beneficial for high-resolution tasks and more stable training compared to FCGAN. For example, pooling operations are replaced with strided convolution and fractional-strided convolution for the discriminator and generator, respectively, batch normalization is used throughout the framework to help locate origin-centered false and generated samples, and ReLU activation with Tanh is used in the generator and discriminator to prevent the entire network from crashing. Arjovsky et al. proposed the Wasserstein GAN [37], which uses the Earth mover (EM) distance as an optimized loss metric instead of the original loss function to improve the stability of training and prevent model collapse. An important difference between the WGAN and the original GAN is that the discriminator in the WGAN is designed to fit the EM distance to help the training converge in a regression manner. The disadvantage of the WGAN is that that suffers from the K-Lipschitz limitation due to weight clipping, and fine-tuning the clipping parameters is difficult. An improved WGAN [38] was proposed to address this drawback. The penalty gradient parametrization is used to force the critic's gradient penalty parametrization around K relative to its output. The results show significantly better performance than the original WGAN and can be stably trained on different GAN frameworks.

The boundary equilibrium GAN [39] combines an equilibrium execution method and a loss derived from EM distance for a GAN with autoencoders to make G and D reach equilibrium during network training. It also provides an approximate convergence determination to decide whether the model has reached its final state. Progressive GAN [40] proposes the idea of progressively overlaying network layers to refine the modeling details and apply them to G and D. This approach is faster in terms of training speed and more efficient in terms of the number of layers. This approach accelerates the training speed while making it well stable. The self-attention mechanism was applied to GAN, resulting in self-attention GAN [41], which can build attention-driven, long-range dependency modeling. It can generate detailed samples using all feature cues from multiple levels. Since improper conditioning of G can lead to degradation of GAN performance, spectral normalization was applied to G to optimize the training process.

Since supervised learning is widely used in CNNs, it has also drawn attention to the GAN framework. Conditional GAN [42] was introduced to conditionally restrict for G and D by inputting real-valued labels. CGAN is fed with additionally labeled data y and is encoded normally before being connected to encoded information z and x. Since CGAN can generate descriptive samples that do not belong to the training labels, several methods combine CGAN loss and L1 or L2 distance between the predicted labels and the ground truth labels. For example, Xu et al. [43] proposed a deep learning method based on the CGAN paradigm to estimate fetal pose from MR volumes. Ji et.al [44] performed saliency detection by using CGAN, which converts saliency map prediction to saliency image segmentation task by using paired images to ground truth saliency maps. There are also some methods using CGAN for semantic segmentation [45], image inpainting [46], and image translation [47]. CGAN shows good performance in these above methods, especially for the task of generating heatmap labels stacked. Thus, we tried to use CGAN in our adversarial framework to improve its performance for the human pose recognition task.

3. Method

Our proposed adversarial framework model, which is elucidated in Figure 2, consists of a pose generator and a pose discriminator. The pose generator network is a convolutional network with an encoding-decoding structure and a multiscale representation. The input to the generator network is a 256×256-sized image with three channels and generates a set of confidence heatmaps that represent the confidence score for the position of each joint. The parameters of the generator are updated by its own forward and backward propagation and by the loss of the discriminator, which is trained in an adversarial manner so that the prior information of the human structure is implicitly utilized.

The details of the individual network architecture are presented in Figure 3. This network architecture combines the improvement in multilevel feature representation [20] and the coding-decoding structure, which combines a stacked hourglass network [21] and a multiscale feature representation with a dilated space pooling module [48]. The multiscale feature representation increases the ability of the network to capture contextual information and can provide sufficient local features for the confidence heatmap of joints and KBOF detection, which can effectively improve performance. Its details are shown in Section 3.1.

3.1. Multiscale Feature Representation for Joint Heatmaps and KBOF. The multiscale feature representation module is shown in Figure 4 for the details of the network structure and task pipeline. Based on the modified HRNet [20] followed by two consecutive head networks, the front head subnet predicts a set of confidence heatmaps K corresponding to each joint point of the body part, and the back head

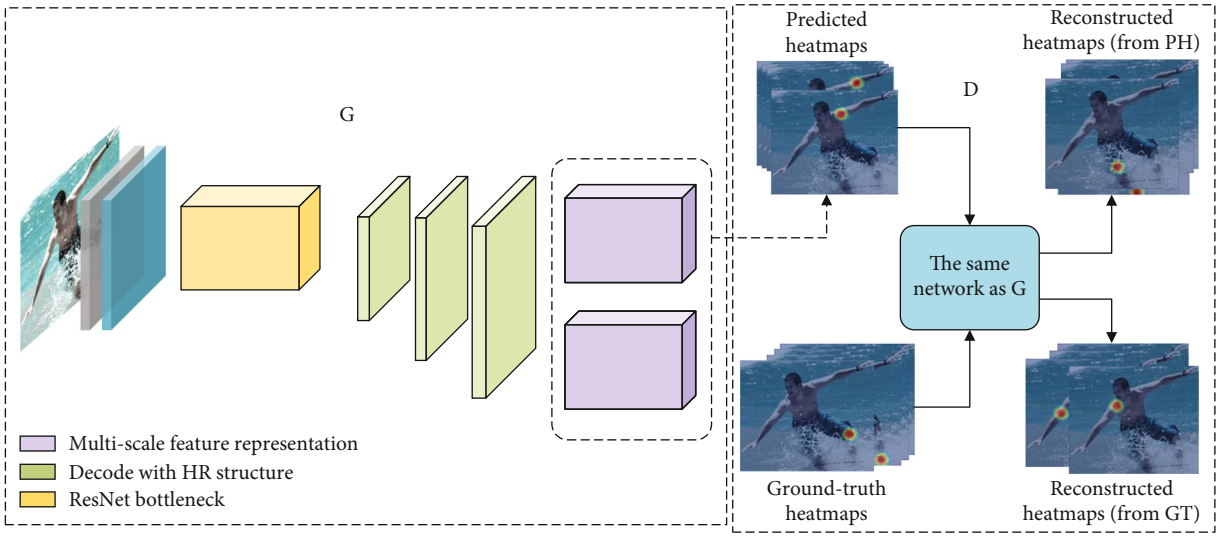


Figure 2: The overall pipeline of our adversarial framework. We utilize a modified HRNet network with a core component, i.e., multiscale feature representation for keypoint heatmap and KBOF detection, while a discriminator is used to identify the generated predicted heatmaps from the ground truth heatmaps by reconstructing the corresponding heatmaps separately. The generators and discriminators have the same network architecture.

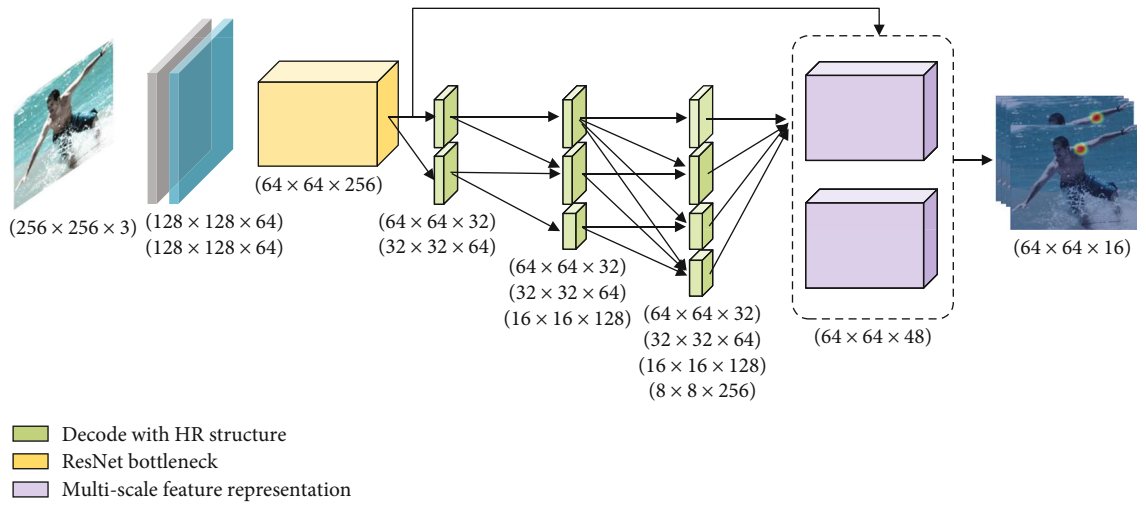


Figure 3: Presentation of the generator framework. We highlight the multiscale feature representation used for joint confidence heatmap and KBOF detection in the purple square. The more specific details of this section are presented in Figure 4.

subnet predicts a set of 2D vector fields L of keypoint bidirected fields (KBOF) which encode bidirectional position and direction information between keypoints of the body. The information between two keypoints is abstractly represented as a strongly connected graph containing two points through the keypoint bidirectional field. The set $K = \{K_1, K_2, \cdots, K_m\}$ contains M keypoint confidence heatmaps, $H_m \epsilon \mathbb{R}^{w \times h}$, $m \epsilon \{1, 2, \cdots, M\}$, and each body joint point corresponds to a heatmap. The set $L = \{L_1, L_2, \cdots, L_n\}$ contains N vector field pairs, where $L_n = \{L_{n1}, L_{n2}\}$, $L_n \epsilon \mathbb{R}^{w \times h \times 2}$, $n \epsilon \{1, 2, \cdots, N\}$ corresponding to a pair of vector fields between each two keypoints. L_{n1} and L_{n2} denote the position and orientation information between keypoints K_i to K_j, $\{i, j\} \epsilon \{1, 2, \cdots, M\}$. The M keypoint confidence heatmaps and the N pairs of bidirectional fields together jointly generate the final predicted heatmaps. Among them, the N pairs of KBOF play an important role in guiding the generation of the final predicted heatmap, which ultimately leads to the generation of predicted keypoints close to the ground truth heatmap.

3.1.1. Network Architecture. The multiscale feature representation module we use, as shown in Figure 4, produces multiscale feature representations for KBOF detection, shown in light orange, and confidence maps for joints, shown in blue. Among them, the guide field is predicted in two consecutive multiscale representation stages, with low-level feature fusion. It increases the perceptual field of the network to have consistent high-resolution processing of the feature representation, which contributes to high regression rates. The multiscale module is based on the dilated convolution

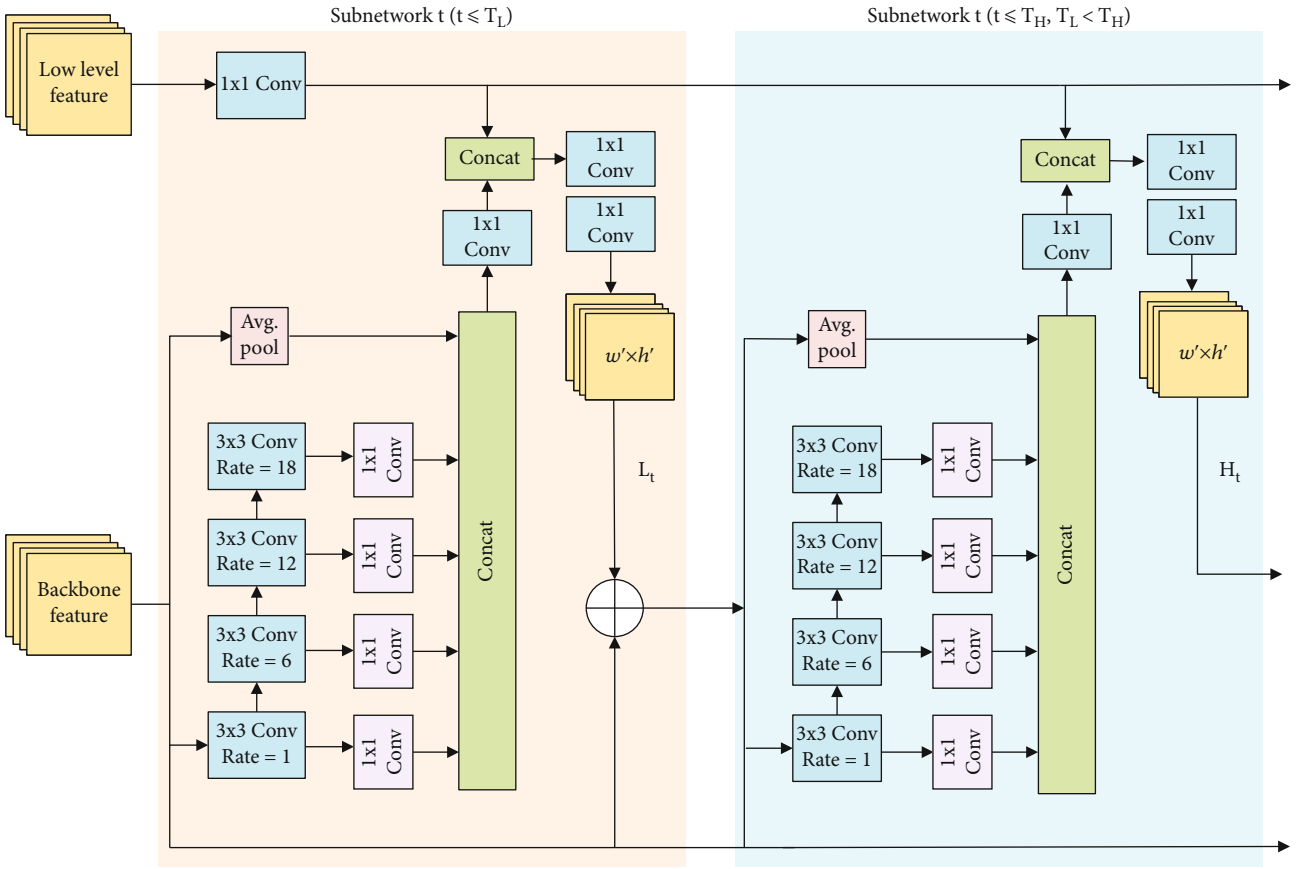


FIGURE 4: Multiscale feature representation for KBOF and joint confidence map detection. The first set of consecutive stage prediction KBOF sets L follows a series of stage prediction joint confidence map sets K. Each stage prediction and its corresponding original image features consist of backbone and low level.

to maintain the multiscale perceptual field and performs a series of parallel inflation convolutions with different inflation rates to increase efficiency. In addition, each parallel module has a parallel averaging pooling layer for the original scale features to increase the overall feature representation.

The multiscale representation module is designed to increase the representation of multiscale features and reduce the number of parameters to solve the memory limitation and overcome the drawbacks of inflationary convolution. His four branches have different expansion rates with ratios of 1, 6, 12, and 18, respectively. This module combines the decoder in the integration unit and processes the four branches and the low-level features at the same higher resolution, leading to more accurate node prediction. The output feature f_{subnet} for each head subnet is described as follows.

$$f_{\text{branch}} = K_1 * \left(\sum_{i=1}^{4} \left(\left(K_{a_i} * f_{i-1} \right) * K_1 \right) + \text{AP}(f_0) \right),$$
$$f_{\text{subnet}} = K_1 * \left(K_1 * \left(K_1 * f_{lf} + f_{\text{branch}} \right) \right), \quad (1)$$

where $*$ represents the convolution, f_{lf} represents the low-level feature map, f_0 represents the input original feature map, f_{i-1} represents the $(i-1)$st feature map from the inflated convolution, K_1 represents the convolution with a convolution kernel size of 1×1, and K_{a_i} represents the convolution kernel with a 3×3 dilated convolution whose expansion rates are $a_i = \{1, 6, 12, 18\}$, as shown in Figure 4.

The multiscale feature representation combines feature representations from lower levels in a connected manner, and the final two layers of 1×1 convolution recover the number of feature maps to the number of joints corresponding to the human pose estimation. Since the high resolution is applied to the backbone of the modified HRNet-based network, this module can directly output the confidence heatmap used to connect the next layer without additional decoding networks or linear interpolation networks to recover its size.

3.1.2. Confidence Heatmap of Keypoint and KBOF Detection. A series of feature maps, which are fed into the first subnet, are generated through the HRNet network with enhancements. A series of human keypoint bidirected fields are generated through this subnet. In the subsequent subnetworks, the predicted feature maps from the previous subnetwork and its combination of feature maps f_{lf} from the low-level network and the original image feature maps f are

concatenated, and further refined predicted feature maps are generated.

$$\begin{cases} L_1 = \varphi^1(f), t = 1, \\ L_t = \varphi^t(L_{t-1}, f), 2 \le t \le T_L, \end{cases} \tag{2}$$

where φ^t refers to the KBOF subnet for inference at subnet t, f is the original image feature map, L_t is the series of generated KBOF through φ^t, and T_L is the total number of subnets of KBOF. After T_L subnets, the confidence map of the node locations is generated through the subsequent subnets with the latest KBOF feature maps.

$$\begin{cases} H_t = \omega^t\left(f, L_{T_L}\right), t = T_L, \\ H_t = \omega^t\left(f, L_{T_L}, H_{t-1}\right), T_L \le t \le T_H, \end{cases} \tag{3}$$

where ω^t refers to the subnet t that generates the confidence graph, L_{T_L} is the latest KBOF generated through ω^t, H_t is the confidence graph generated, and T_H is the total number of subnets in this module including the subnets that generate the KBOF.

We empirically traded off the number of subnets of KBOF to improve the prediction results of the confidence map, which has two detection subnets and one detection subnet for the nodal heatmap because the multiscale feature representation no longer requires more redundant subnets. The multiscale feature representation can be processed in parallel and extracted by multiscale inflation convolution. Thus, the computational effort of each subnet and the parameters of the overall network are reduced. To guide the network to refine the prediction of correlation information for KBOF in the first stage and confidence maps in the second stage, we added a loss function at the end of each stage. The probability of generating anomalous sample points is relatively small because of the structural limitations of the human body. So L_2 loss is used between the ground truth and the predicted value. The loss function of the KBOF branch in the subnet t_i and the loss function of the Gaussian graph branch in the subnet t_j are as follows:

$$\mathscr{L}_L^{t_i} = \frac{1}{NP}\sum_{n=1}^{N}\sum_{p}\left\|L_n^*(p) - L_n^{t_i}(p)\right\|_2^2, \tag{4}$$

$$\mathscr{L}_H^{t_j} = \frac{1}{MP}\sum_{m=1}^{M}\sum_{p}\left\|H_m^*(p) - H_m^{t_j}(p)\right\|_2^2, \tag{5}$$

where L_n^* refers to the ground truth KBOF, $L_n^{t_i}$ is the predicted KBOF in the subnetwork t_i, H_m^* is the ground truth confidence map, $H_m^{t_j}$ is the predicted confidence map, M refers to the total number of keypoints of the human joints, and N refers to the total number of bidirectional oriented field pairs. The gradient vanishing problem is solved by adding intermediate supervision at the end of each subnet. The overall loss in multiple subnets can be expressed as

$$\mathscr{L} = \sum_{t=1}^{T_L}\mathscr{L}_L^t + \sum_{t=1}^{T_L+T_H}\mathscr{L}_H^t. \tag{6}$$

3.1.3. Confidence Maps for Keypoint Detection. Each confidence map, referred to as a Gaussian heatmap, is a 2D representation of the belief that a body part is designated at any location in the picture. To evaluate the loss function $\mathscr{L}_H$ in equation (5), ground truth confidence maps H^* are generated from the 2D keypoints that are labeled. Confidence map $H_{m,k}^*$ is generated for a single person k. Note that $x_{m,k}$ is the true position of body part m of individual k in the picture. Then, for $H_{m,k}^*$, the value at p pixels is defined as

$$H_{m,k}^*(p) = \exp\left(-\frac{\left\|x_{m,k} - p\right\|_2^2}{\sigma^2}\right), \tag{7}$$

where σ controls the expansion of the vertices. A value of σ of 3 was empirically obtained in this experiment, resulting in well-defined Gaussian curves both for the ground truth heatmap and for the predicted output heatmap.

3.1.4. Keypoint Biorientation Field for Keypoint Association. The human keypoint bidirectional oriented field contains both position- and orientation-oriented information in the limb support region. Inspired by [8], we propose that each KBOF contains a pair of 2D vector fields for each limb, as shown in Figure 1(c). For each pixel in the region belonging to a particular limb, a pair of 2D vector fields encodes the orientation, i.e., fr5om one joint point of the limb to another joint point. Each type of limb has a corresponding pair of KBOFs connecting its associated body keypoints.

Consider that each limb part comes from the human body (see Supplementary Materials (available here)). Denote $x_{n,k}$ and $y_{n,k}$ as the ground truth values of the two keypoints of the body part from limb n for person k in the picture, respectively. If two outside a point p lies within the limb, $L_{n,k}^*(p)$ is a pair of unit vectors at pixel point p, $L_{n,k,1}^*(p)$ and $L_{n,k,2}^*(p)$, respectively. The value of $L_{n,k,1}^*(p)$ is a unit vector from point x to y, and the value of $L_{n,k,2}^*(p)$ is an inverse unit vector. The value of points located outside the limb is 0.

To evaluate the loss function $\mathscr{L}_L$ in equation (5), the ground truth value L^* of KBOF is generated at point p in the image. The value of $L_{n,k,1}^*$ at pixel p for a limb n of person k is defined as

$$L_{n,k,1}^*(p) = \begin{cases} \dfrac{\left(x_{n,k} - y_{n,k}\right)}{\left\|x_{n,k} - y_{n,k}\right\|_2}, & \text{if } p \text{ on limb } n, k, \text{ from } x \text{ to } y, \\ 0, & \text{otherwise}, \end{cases}$$

$$L_{n,k,2}^*(p) = \begin{cases} \dfrac{\left(y_{n,k} - x_{n,k}\right)}{\left\|x_{n,k} - y_{n,k}\right\|_2}, & \text{if } p \text{ on limb } n, k, \text{ from } y \text{ to } x, \\ 0, & \text{otherwise}. \end{cases} \tag{8}$$

The set of points on a limb is defined as a point within a distance threshold based on a line segment as follows:

For $L^*_{n,k,1}(p)$, the range of p is such that

$$\begin{aligned} 0 \le u \bullet (p - x_{n,k}) \le l_{n,k}, \\ \left| v \cdot (p - x_{n,k}) \right| \le b_l. \end{aligned} \tag{9}$$

For $L^*_{n,k,2}(p)$, the range of p is such that

$$\begin{aligned} 0 \le u \bullet \left(p - y_{n,k}\right) \le l_{n,k}, \\ \left| v \cdot \left(p - y_{n,k}\right) \right| \le b_l, \end{aligned} \tag{10}$$

where b_l refers to the distance of the limb width in pixel points, the limb length is $l_{n,k} = \|x_{n,k} - y_{n,k}\|_2$, u is the unit vector parallel to the limb direction, and v is the unit vector perpendicular to u.

We measure the degree of association between the predicted heatmaps by computing line integrals over the corresponding KBOF along the line segments connecting the keypoints. In particular, for predicted nodes k_1 and k_2, portions of the predicted KBOF are sampled and L_n along the sampled line segments is employed to measure the confidence level of the association between them.

$$S = \int_{c=0}^{c=1} L_n(w(c)) \bullet \frac{k_2 - k_1}{\|k_2 - k_1\|_2} dc, \tag{11}$$

where $w(c)$ is a single linear interpolation of the two predicted nodes k_1 and k_2, which is referred to as

$$w(c) = (1 - c)k_1 + ck_2. \tag{12}$$

3.2. Generator. We used a deep CNN architecture based on a modified HRNet as the backbone network of the generator. The processing pipeline of this architecture is shown in Figure 3. Inspired by [20], the idea of HRNet is applied to the deconvolution operation of the backbone network. In addition, we improved the results of the backbone network by using Gaussian modulated deconvolution instead of the upsampling operation. The improved HRNet feature subnetwork is followed by the multitask prediction module we used for the Gaussian heatmap and KBOF. Details of the multitask prediction module have been shown in Section 3.1.

The task of the generator is to learn a mapping that attempts to project a color picture x onto its corresponding final predicted heatmap y. The predicted heatmap y is generated by the joint guidance of the native predicted heatmap and KBOF as described in Section 3.1. The generator generates M final key heatmaps y (16 and 17 in the datasets MPII and COCO, respectively), each of which is a 64×64 mapping having a Gaussian heatmap at each joint location. Combined with the detailed network LOSS described in Section 3.1.2, the average mean square error LOSS function of our proposed generator network can be described as

$$\mathscr{L}_{\text{Mse}} = \sum_{m=1}^{M} \|y^*_m - y_m\|_2^2, \tag{13}$$

where y^*_m is the ground truth heatmap, y_m is the final predicted heatmap of the generator, and M is the total number of human joints. By backpropagating the $\mathscr{L}_G(\theta)$ gradient, the generator is forced to learn the image features based on the keypoint heatmap and KBOF to further improve the accuracy.

In addition, we add the adversarial loss, which can help the generator to generate reasonable poses. The adversarial loss function can be expressed as

$$\mathscr{L}_{\text{Adv}} = \sum_{m=1}^{M} \|y_m - D(y_m, x)\|_2^2, \tag{14}$$

where D is the discriminator and x is the original input image. $\mathscr{L}_{\text{Adv}}$ measures the error between the predicted heatmap of the generator and the reconstructed heatmap of the discriminator.

In summary, the joint loss function of the generator can be expressed as

$$\mathscr{L}_G = \mathscr{L}_{\text{Mse}} + \alpha_G \mathscr{L}_{\text{Adv}}, \tag{15}$$

where α_G is the hyperparameter that controls the weight of the $\mathscr{L}_{\text{Adv}}$ and $\mathscr{L}_{\text{Mse}}$.

3.3. Discriminator. The generator network learns the association between local keypoints starting from the connection between local features. These benefits from KBOF are guiding the intermediate prediction heatmaps to generate more accurate final keypoint heatmaps on local features. But this is not enough, and discriminators are introduced to help the generator optimize the detection node confidence map from the body joint configuration from the local perspective, i.e., embedding the body joint configuration a priori into the network architecture. The discriminator is used to distinguish reasonable pose predictions from implausible pose estimates, which helps the generator further improve the accuracy of the predictions. The processing pipeline of discriminator is shown in Figure 5.

The final heatmap generated by the generator is fed to the discriminator along with the ground truth heatmap. For each pair of input images, the discriminator should go through and distinguish whether this final heatmap is reasonable or not. Both are reconstructed at the same time to generate the reconstructed final heatmap and the reconstructed ground truth heatmap, similar to the code-and-decode operation. The quality of the reconstructed heatmap depends on how similar it is to the final input heatmap. The error between these two pairs of heatmaps is measured by two loss functions $\mathscr{L}_{\text{Fake}}$ and $\mathscr{L}_{\text{Real}}$, respectively.

For the input ground truth heatmap, the discriminator is trained to extract its features and reconstruct a similar ground truth heatmap. For the final input heatmap, the

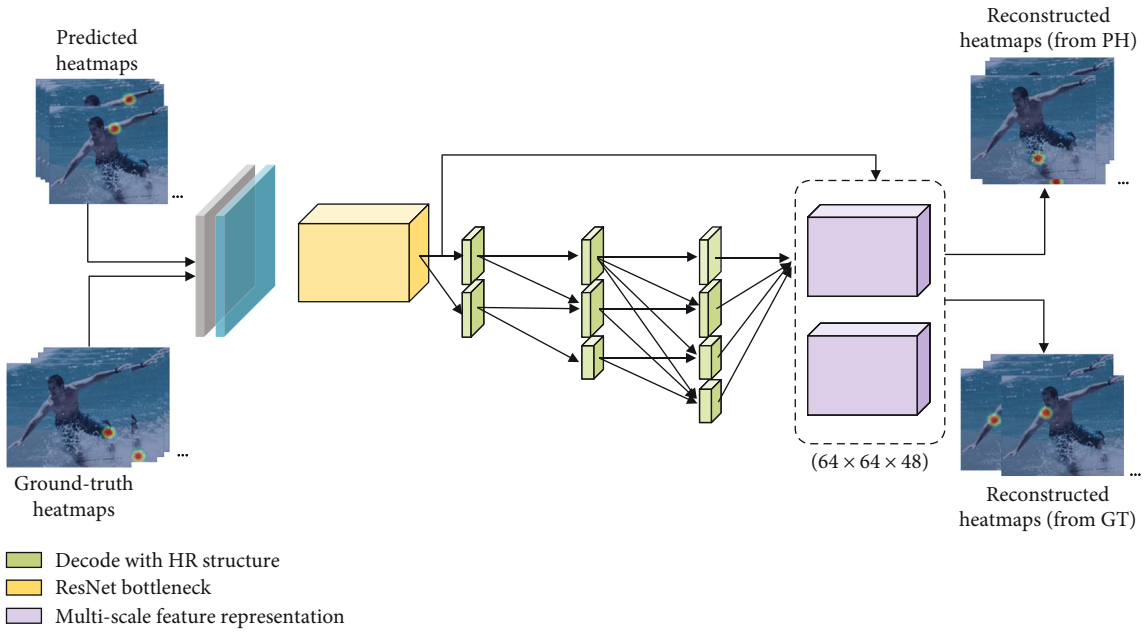


FIGURE 5: Presentation of our discriminator framework. The discriminator is used to distinguish the predicted heatmaps from the ground truth heatmaps by reconstructing both corresponding input heatmaps.

discriminator is trained to generate a completely different heatmap. That is, the error between the input ground truth heatmap and the reconstructed one is minimized, and the error between the input final predicted heatmap and the reconstructed one is maximized. The loss functions $\mathscr{L}_{\text{Fake}}$ and $\mathscr{L}_{\text{Real}}$ are described as follows.

$$\mathscr{L}_{\text{Fake}} = \sum_{m=1}^{M} \|y_m - D(y_m, x)\|_2^2, \tag{16}$$

$$\mathscr{L}_{\text{real}} = \sum_{m=1}^{M} \|y_m^* - D(y_m^*, x)\|_2^2, \tag{17}$$

$$\mathscr{L}_D = \mathscr{L}_{\text{real}} - k_t \mathscr{L}_{\text{Fake}}, \tag{18}$$

where β is a super control variable introduced, which helps the GAN to converge during training, as detailed below. Since the loss function is computed over each pixel point, as described in Section 3.1.2, this makes the discriminator act like a harsh critic, giving a detailed evaluation of the final heatmap of the input and rejecting unreasonable joint heatmaps. This is where it differs from traditional GAN.

As mentioned in many papers [35, 49, 50], GANs are unstable and difficult to train as the discriminator simply give a 0 or 1 evaluation, making the network difficult to converge. Inspired by [39], we used a variable k_t to control the LOSS from the generator and controller, which is updated at each iteration t. The adaptation k_t is described as follows.

$$k_{t+1} = k_t + \beta(\gamma \mathscr{L}_{\text{Real}} - \mathscr{L}_{\text{Fake}}), \tag{19}$$

where k_t has a value between 0 and 1 and β and γ are two hyperparameters. When the performance of the generator is due to the discriminator, i.e., the heatmap generated by the generator can deceive the discriminator, $\mathscr{L}_{\text{Fake}}$ is smaller than $\gamma \mathscr{L}_{\text{Real}}$. k_t becomes larger, making the weight of $\mathscr{L}_{\text{Fake}}$ increase, so the discriminator will be trained to recognize the true pose heatmap from the fake one. $\gamma \mathscr{L}_{\text{Real}} - \mathscr{L}_{\text{Fake}}$ determines the magnitude of the value addition of k_t; i.e., the discriminator's performance lags behind the generator's magnitude, with a larger magnitude adding more value and a smaller magnitude adding less value. Conversely, k_t decreases when the discriminator outperforms the generator, allowing the generator's performance to catch up with the discriminator.

3.4. Adversarial Training. We guide the network to generate more accurate keypoints through the implementation of KBOF based on local features. Also, we applied the idea of the BEGAN [34] network to help the generator further improve the accuracy of the human keypoint task by embedding the human structure a priori into the network.

$\mathscr{L}_{\text{Adv}}$ in equation (14) and $\mathscr{L}_{\text{Fake}}$ in equation (16) have the same value, but the multiplied weights and directions are different. The generator minimizes $\mathscr{L}_{\text{Mse}}$ and $\mathscr{L}_{\text{Adv}}$, and the discriminator maximizes $\mathscr{L}_{\text{Fake}}$ and minimizes $\mathscr{L}_{\text{real}}$. This is the adversarial process that constitutes the architecture of this network. This causes the generator to evolve towards generating high confidence, while the discriminator strives to find features that help generate the correct joint heatmap to distinguish the high confidence heatmap from the low confidence heatmap. Algorithm 1 shows the adversarial training process for this network architecture.

Require: The unlabeled color image x and the corresponding ground truth heatmap $y*$
1 Forward Discriminator by $\{y_{real}\} = \mathscr{D}(x, y^*)$
2 Compute the gradient ∇f_D according to Eq. (17)
3 Forward Generator by $\{y\} = G(x)$
4 Compute the gradient ∇f_G according to Eq. (13)
5 Foreword Discriminator by $\{y_{fake}\} = \mathscr{D}(x, y)$
6 Cumulative gradient ∇f_D for Eq. (16)
7 Update Discriminator with ∇f_D
8 Cumulative gradient ∇f_G for Eq. (14)
9 Update Discriminator with ∇f_G
10 Return to step 1 until the accuracy rate no longer increases

Algorithm 1: The adversarial training process of our framework.

4. Experiments

We conducted experiments on two widely used benchmark datasets, i.e., MPII and COCO datasets, and analyzed the experimental results in detail and comprehensively. In particular, we conducted analytical experiments on the MPII dataset in a masking context. We have performed a quantitative analysis of the experimental results using a generic evaluation while showing the results of the qualitative analysis. The relevant details of the experimental part are also described in this section.

4.1. Datasets. The MPII Human Pose dataset [51] refers to the Max Planck Institute for Informatics Dataset. MPII is the most widely used benchmark for the evaluation of human posture estimation of joints. This dataset contains approximately 25K photographs of over 40K people with labeled body joints. These images were systematically collected using an established classification system for daily human activity. This dataset has over 410 human activities, and each image is provided with an activity label. In addition, the test set is richly annotated with body part occlusion and 3D torso and head orientation. The target annotation contains 16 keypoints.

The COCO dataset [52] refers to the Microsoft Common Objects in Context Dataset. This dataset contains about 200K photos and 250K human instances with keypoint labels, which are divided into a training set, validation set, and test set. In the human label annotations, medium- and large-scale human instances account for most of the instances. This dataset simultaneously predicts the human body and locates 17 keypoints, including 12 human keypoints and 5 facial keypoints for each person.

4.2. Evaluation Metrics. Percentage of Correct Keypoints (PCK) [51] calculates the percentage of the normalized distance between the detected keypoints and their corresponding ground truth heatmaps that are less than a set threshold. Half of the head length is used as the normalization reference in MPII, which refers to PCKh@0.5. It indicates the percentage of the correct proportion of predicted critical points. The equation can be expressed as follows.

$$\mathrm{PCK}_i^k = \frac{\sum_p \delta\left(d_{p^i}/d_p^{\mathrm{def}} \le T_k\right)}{\sum_p 1}, \tag{20}$$

where i denotes the i-th keypoint, p denotes the p-th person, k denotes the k-th threshold T_k, d_{p^i} denotes the Euclidean distance between the predicted value of the i-th keypoint of the p-th person and the ground truth, and d_p^{def} denotes the scale threshold of the p-th person, which in MPII is the head length as the normalized reference.

Object Keypoint Similarity (OKS) [52], in the human keypoint evaluation task, for the goodness of the keypoints obtained by the network, is not computed just by a simple Euclidean distance, but a certain scale is added to compute the similarity between two points. This metric is mainly used in the multiperson pose estimation task. Its equation is as follows:

$$\mathrm{OKS}_p = \frac{\sum_i \exp\left(-d_{p^i}^2/2S_p^2\sigma_i^2\right)\delta\left(v_{p^i} > 0\right)}{\sum_i \delta\left(v_{p^i} > 0\right)}, \tag{21}$$

where p^i denotes the keypoint i of the pth person, d_{p^i} denotes the Euclidean distance between the predicted value of the current key and ground truth, $v_{p^i} = 1$ means that the visibility of this keypoint is 1, $v_{p^i} = 2$ means that this keypoint is occluded but labeled, S_p denotes the scale factor of the pth person of ground truth, and σ_i denotes the keypoint normalization factor. $\delta(*)$ indicates that $\delta(*) = 1$ if the condition $*$ holds; otherwise, $\delta(*) = 0$.

Average Precision (AP) and Average Recall (AR): AP measures the ratio of the predicted keypoint accuracy, i.e., the ratio of true positive samples to the total positive samples. AR measures the ratio of the predicted keypoint regression rate; i.e., the ratio of predicted positive samples AP and AR has many variants to further measure the accuracy of the prediction results, and their detailed description can be found in the literature [52]. AP and AR are often combined with OKS to evaluate the experimental results.

Figure 6: The qualitative results of some sample pictures in the MPII (top) and COCO (bottom) datasets. These scenes include self-occlusion, other people occlusion, and object interference.

Table 1: Results of our framework and comparison with SOTA methods on the MPII test set (PCKh@0.5).

Method	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Total
Carreira et al. [58]	95.7	91.7	81.7	72.4	82.8	73.2	66.4	81.3
Hu & Ramanan. [59]	95.0	91.6	83.0	76.6	81.9	74.5	69.5	82.4
Newell et al. [21]	98.2	96.3	91.2	87.1	90.1	87.4	83.6	90.9
Yang et al. [60]	98.5	96.7	92.5	88.7	91.1	88.6	86.0	92.0
Ke et al. [61]	98.5	96.8	92.7	88.4	90.6	89.4	86.3	92.1
Tang et al. [62]	98.4	96.9	92.6	88.7	91.8	89.4	86.2	92.3
Sekii [63]	97.9	95.3	89.1	83.5	87.9	82.7	76.2	88.1
Sun et al. [20]	97.1	95.9	90.3	86.5	89.1	87.1	83.3	90.3
Baseline [57]	98.5	96.6	91.9	87.6	91.1	88.1	84.1	91.5
Zhang et al. [54]	97.2	95.9	91.2	86.7	89.7	86.7	84.0	90.6
Zhang et al. [64]	98.3	96.4	91.5	87.4	90.9	87.1	83.7	91.1
Artacho et al. [3]	—	—	—	—	—	—	—	92.7
Tang et al. [51]	98.7	97.1	93.1	89.4	91.9	90.1	86.7	92.7
Ours(-)[1]	97.9	95.9	91.3	88.3	91.2	89.7	86.1	91.5
Ours (KBOF+GAN)	98.7	97.3	94.2	91.1	93.0	92.7	88.7	93.9

Table 2: Results of our framework and comparison with SOTA methods on the MPII full test set (mAP@0.5).

Method	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	mAP
Iqbal & Gall [66]	58.4	53.9	44.5	35.0	42.2	36.7	31.1	43.1
Levinkov et al. [67]	89.8	85.2	71.8	59.6	71.1	63.0	53.5	70.6
Insafutdinov et al. [68]	88.8	87.0	75.9	64.9	74.2	68.8	60.5	74.3
Cao et al. [8]	91.2	87.6	77.7	66.8	75.4	68.9	61.7	75.6
Newell et al. [21]	92.1	89.3	78.9	69.8	76.2	71.6	64.7	77.5
Fieraru et al. [65]	91.8	89.5	80.4	69.6	77.3	71.7	65.5	78.0
Our	91.6	87.9	81.6	72.5	77.5	73.8	67.4	78.9

4.3. Experimental Settings. The batch size depends on the size of the input dataset images. Multiscale feature representation module using dilated convolution has different expansion rates which are available. We found in our experiments that a too large expansion rate leads to a decrease in accuracy performance. This is because a too large expansion rate causes the network to not fully capture the correlation information between detailed local keypoints in the image. Therefore, a range of expansion rates $r = \{1, 6, 12, 18\}$ was chosen for the dilated convolution module.

We followed the data augmentation in [53]. We randomly flip and input data and randomly scale with

Table 3: Results of our framework and comparison with SOTA methods on the COCO set for validation.

Method	Input size	PARAMs	GFLOPs	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
CPN [69]	384×288	—	—	72.2	89.2	78.6	68.1	79.3	—
Baseline [57]	384×288	68.6	35.6	74.3	89.6	81.1	70.5	81.6	79.7
EvoPose2D [51]	384×288	7.3	5.6	75.1	90.2	81.9	71.5	81.7	81.0
HRNet [69]	384×288	63.6	32.9	76.3	90.8	82.9	72.3	83.4	81.2
DarkPose [54]	384×288	63.8	34.5	76.8	90.6	83.2	72.8	84.0	81.7
Our	384×288	74.8	31.2	78.1	92.3	85.9	75.6	83.9	82.0

Table 4: Results of our framework and comparison with SOTA methods on the COCO set for test-dev.

Method	Input size	PARAMs	GFLOPs	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
G-RMI [70]	353×257	—	57.0	64.9	85.5	71.3	62.3	70.0	69.7
CPN [69]	384×288	—	—	72.1	91.4	80.0	68.7	77.2	78.5
Cai et al. [58]	256×192	8.7	6.4	72.5	93.0	81.3	69.9	76.5	78.8
Baseline [57]	384×288	68.6	35.6	73.7	91.9	80.0	68.7	77.2	78.5
HRNet [20]	384×288	63.6	32.9	75.5	92.5	83.3	71.9	81.5	80.5
MSPN [25]	384×288	120	19.9	76.1	93.4	83.8	72.3	81.5	81.6
DarkPose [54]	384×288	63.6	32.9	76.2	92.5	83.6	72.5	82.4	81.1
Our	384×288	69.7	31.2	76.5	92.7	83.5	72.6	82.8	81.4

Table 5: Comparison experiments of the number of subnets for KBOF and confidence maps on the COCO validation. CM refers to the confidence maps for joint, and the number indicates the estimated number of subnets for KBOF and CM. Subnets refer to the total number of KBOF and CM subnets. The smaller the total number of subnets, the more it can increase the runtime performance and, at the same time, affect the index performance. There is a trade-off between running time and index performance.

Method	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR	Subnets
3 KBOF & 1 CM	78.2	92.3	86.0	75.8	83.9	81.9	4
2 KBOF & 2 CM	78.0	92.3	85.9	75.6	83.8	81.8	4
1 KBOF & 3 CM	77.6	91.8	85.4	75.1	83.3	81.6	4
2 CM & 2 KBOF	75.3	91.0	82.4	73.0	80.7	79.8	4
1 KBOF & 2 CM	77.5	91.6	85.3	75.0	83.2	81.6	3
2 KBOF & 1 CM	78.1	92.3	85.9	75.6	83.9	82.0	3

coefficients in $[0.65, 1.35]$ by arbitrarily rotating its data at the angle of $[-45, 45]$. The above approach makes this network more robust for images with different orientations and scales. We used the RMSprop algorithm to train the network and calculated the learning rate of the network based on a stepwise approach where the learning rate was initialized to 2.0×10^{-4}. The model was trained for 250 epochs on the two datasets mentioned in this paper. The learning rate was reduced by order of magnitude at three steps at 170 epochs, 200 epochs, and 220 epochs, respectively, following the procedure set to pass [54]. The experiment was conducted on an NVIDIA Tesla V100 GPU.

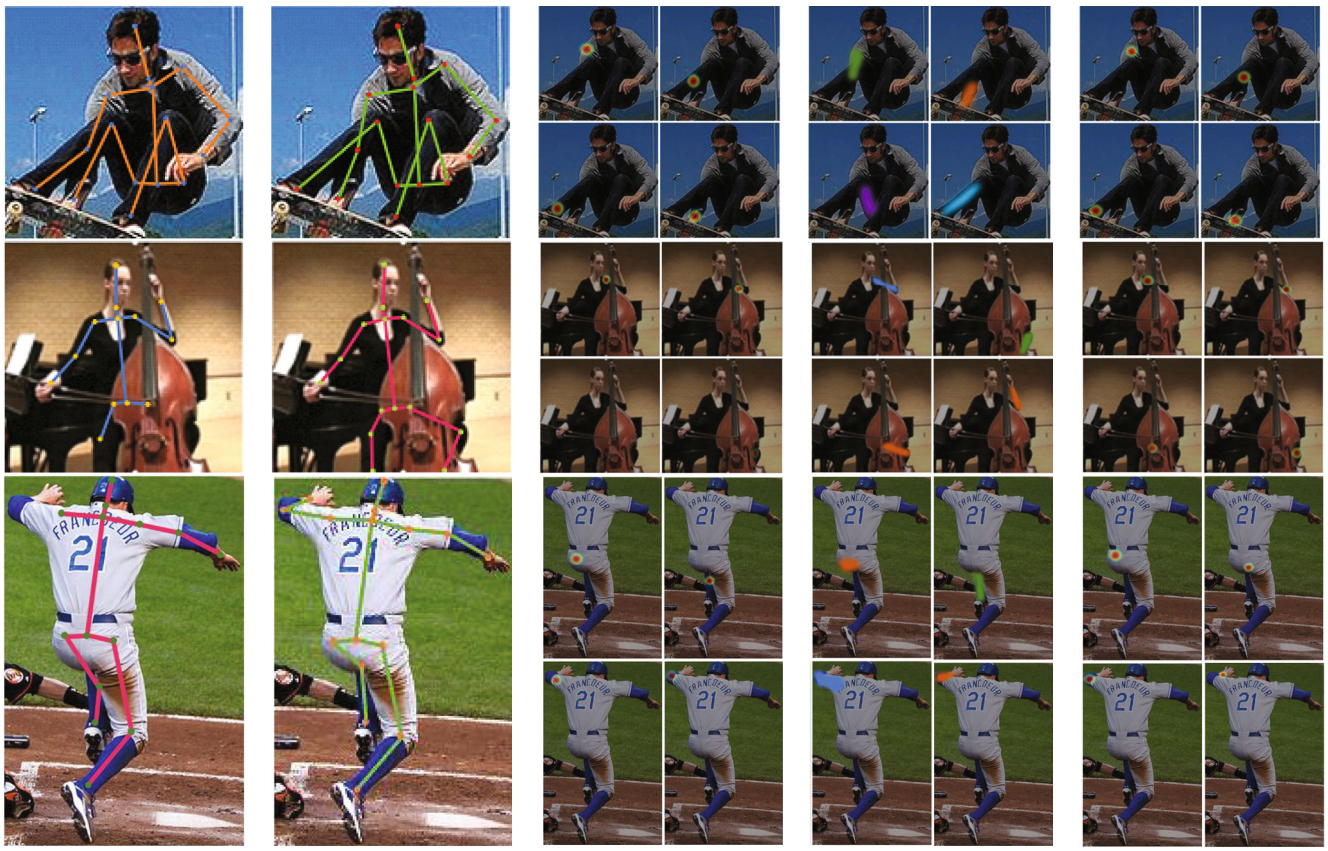
5. Result and Discussion

Our framework is tested on two large and widely used datasets, and the latest methods are compared. And ablation experiments are conducted to detect the contribution of our proposed multitasking module with KBOF guidance, the use of multiscale null convolution, and the conditional generation adversarial framework to improve the accuracy.

5.1. Results on the MPII Dataset. We conducted experiments on the MPII dataset with PCKh@0.5 as the standard measure. Our methods were trained on the COCO training set and fine-tuned on the MPII training set, and data augmentation was performed, which was routinely operated to follow [8, 55, 56]. The results of this framework are presented in Figure 6 and Table 1.

As shown in Table 1, our experimental results achieve a performance of 93.9% and exceed 1.3% relative to the SOTA method. The improvement of 2.6% compared to baseline [57] presents a meaningful improvement relative to the previous SOTA. Our experimental results have a meaningful improvement relative to the previous SOTA. The results of our experiment have a wide range of improvements over each node, especially for the more connected but difficult to accurately detect parts of the body, such as the wrist and ankle. This shows the robustness of our method for different detection sites, especially for the more connected parts of the body. Qualitative results are shown in Figure 6 with successful detection. These examples illustrate that our method can effectively cope with the occlusion problem.

Table 2 shows the comparison between the results of our experiment and the SOTA method with mAP@0.5 as the



(a) (b) (c) (d) (e)

Figure 7: Qualitative result analysis and joint heatmaps and KBOF visualization of our method. (a) The prediction heatmaps generated by the native HRNet [20] are easily disturbed. (b) Our method further refines the heatmaps and produces the most accurate prediction. (c) The initial prediction heatmaps in our network without KBOF. (d) The visualization of KBOF contains the orientation and position information between the two joints to guide the network to further refine the prediction of the joint point heatmaps. (e) The final generated prediction heatmaps are guided by KBOF on the original prediction heatmaps.

measurement scale. As presented in Table 1, the experimental results of this framework have a comprehensive improvement at the joints with strong correlations. Our results are improved by 1.2% relative to the method presented in Fieraru et al. [65].

5.2. Results on the COCO Dataset. We conducted training and testing on the MPII dataset and analyzed PARAM and GFLOPs. Tables 3 and 4 show the training results of our framework on the MPII validation set and the development test set, respectively. This modified version of HRNet, combined with KBOF and the adversarial framework, results in prediction accuracy of 78.1%. This shows that our framework has a significant improvement in the average accuracy. This result is a 2.4% improvement over the original HRNet and a 3.8% improvement over baseline. There is also a 1.7% improvement in prediction accuracy compared to the previous SOTA approach.

Our results present an improvement in prediction accuracy on multiple measures. More significantly, the detection of AP for medium objects in very difficult ones obtained by our method presents an improvement of 3.9% relative to the SOTA method. This demonstrates the improved capability of our framework for predicting difficult poses using a relatively small number of pixel points. This is due to our use of multiscale null convolution combined with KBOF to explore multiple-scale features on the picture and generate plausible human poses by adversarial means.

Qualitative results are presented in Figure 6 for the COCO dataset. From these examples, we can see that our method is robust and has high accuracy for complex scenes. Challenging scenarios include the detection of nodes where limbs are obscured or not sufficiently separated or intertwined, but our method maintains high performance. The experimental results of our method are also presented in the COCO dataset for test-dev. As shown in Table 4, the results also show the high performance of our framework compared to the SOTA method.

5.3. Comparison Analysis of the Number of Subnets. In Table 5, we analyze the impact of KBOF refinement on the final confidence map predictions generated. We fixed the total number of subnets as 3 and 4 subnets, respectively, in which KBOF and confidence map have different numbers

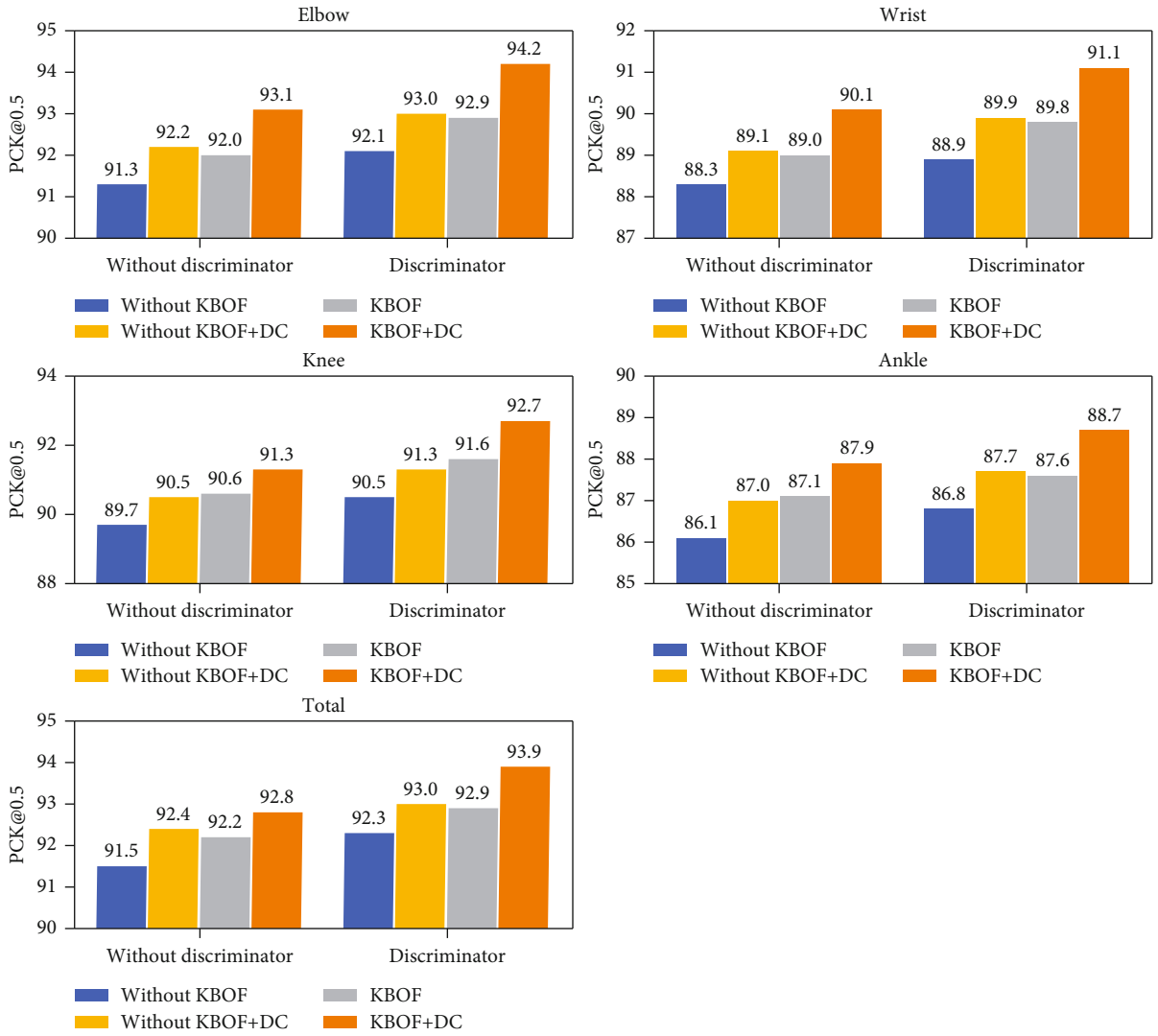


Figure 8: Several different versions of our framework for ablation experiments. KBOF represents the use of Keypoint Biorientation Fields, MSDC denotes the use of multiscale dilated convolution module for KBOF and part heatmap detection, and the discriminator denotes the use of the adversarial framework in this network.

of subnets. We can get three conclusions from this experiment. First, because the multiscale feature extraction structure is applied to the network, the total number of three subnets reaches almost the same performance index as the number of four subnets. This is because the features have been fully extracted in our network. To make a trade-off between running time and performance index, we finally fixed the total number of subnets to 4. Second, KBOF requires more subnets to refine compared to confidence maps. Third, the accuracy of the final confidence map will be dramatically increased when using KBOF as the forerunner, but vice versa will result in a 2.7% reduction for absolute accuracy. Even if the model has a total number of subnets of 4 (2 CM and 2 KBOF), the model with less accuracy than the total number of 3 subnets with higher computational cost predicts KBOF (2 KBOF and 1 CM) first.

Meanwhile, qualitative results with joint heatmaps and KBOF visualization of our method are presented in Figure 7, which are compared with the inference results of HRNet. It can be seen that after KBOF's guidance, the accuracy of the joint heatmap has been improved.

5.4. Ablation Experiment. Ablation experiments were performed on the MPII dataset to explore the effectiveness of several different aspects of our approach. A modified HRNet-based native network was used as the baseline, as shown in "Ours(-)[1]" in Table 1. Several key results of the experiments are presented in Figure 8. We conducted analytical experiments for three parts used in our framework: KBOF, the use of multiscale convolution, and discriminators.

As shown in the total section in Figure 8, we first removed the discriminator and performed an exploratory study on the use of KBOF and null convolution. The accuracy of the prediction results is improved by about 0.8% by the use of KBOF. This shows that KBOF can guide the

network to generate predictive joint heatmaps with higher accuracy. And the multiscale cavity convolution results in a contribution of about 0.6 relative to "Ours(-)[1]." This also indicates that the multiscale cavity convolution further improves the network's ability to extract valid features. Finally, on top of the first two, the discriminator is introduced, resulting in a boost of about 0.7. The discriminator can exclude the seemingly unreasonable poses generated by the generator, thus allowing the generator to generate seemingly reasonable predictions of human poses. The generator was used to generate the final heatmap by removing the discriminator, which resulted in an accuracy improvement.

Overall, the addition of isolated individual components was all able to increase the accuracy of the predicted heatmap. But using these components separately resulted in boosts of about 0.8, 0.6, and 0.7, respectively, and the joint use of these components produced a 2.4% boost. This is probably because the null convolution extracts valid features, and KBOF further makes a boost on top of these valid features. The discriminator can help the generator to fully recognize the heatmap features and guide it to generate seemingly reasonable predictions of the human pose by exploiting the valid results from the previous step.

6. Conclusion

We propose a novel adversarial framework with the Keypoint Biorientation Field (KBOF) for multiperson pose estimation. The present nonparametric representation can effectively encode the orientation and location information between the nodes, forming a strong connectivity graph similar to the one containing two nodes (two nodes can reason about each other's information), which can guide the network to generate a more accurate heatmap of the nodes. In addition, multiscale null convolution is used to extract effective local features in the image for KBOF. Finally, a discriminator with the same network structure as the generator is introduced to embed the human joint prior into the network to further improve the accuracy of the predicted joint point heatmap. The discriminator is removed after the framework is trained, and we only need to use the generator for testing. There is no additional computational overhead, so it does not affect the inference time of the actual network. The bidirectional nonparametric representation presented in this method can be further developed and evolved for application to tasks with structural information, such as 3D human pose detection, face landmark detection, and finger movement detection.

Data Availability

The code and data used to support the research are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this article.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (62702416) and the National Key Research and Development Program of China (No. 2019YFC1521102).

Supplementary Materials

Figure S1: a pair of unit vectors has opposite directions between the joint points. *(Supplementary Materials)*

References

[1] Y. Yu, H. Li, J. Cao, and X. Luo, "Three-dimensional working pose estimation in industrial scenarios with monocular camera," *IEEE Internet of Things Journal*, vol. 8, no. 3, pp. 1740–1748, 2020.

[2] X. Li and D. Li, "GPFS: a graph-based human pose forecasting system for smart home with online learning," *ACM Transactions on Sensor Networks (TOSN)*, vol. 17, no. 3, pp. 1–19, 2021.

[3] B. Artacho and A. Savakis, "Unipose: unified human pose estimation in single images and videos," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, June 2020.

[4] W. Tang and Y. Wu, "Does learning specific features for related parts help human pose estimation?," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, 2019.

[5] Y. Cai, Y. Cai, Z. Wang et al., "Learning delicate local representations for multi-person pose estimation," in *European Conference on Computer Vision*, Springer, 2020.

[6] Y. Chen, Y. Chen, C. Shen, X.-S. Wei, L. Liu, and J. Yang, "Adversarial posenet: a structure-aware convolutional network for human pose estimation," in *Proceedings of the IEEE International Conference on Computer Vision*, Venice, Italy, 2017.

[7] Z. Cao, Z. Cao, R. Wang, X. Wang, Z. Liu, and X. Zhu, "Improving human pose estimation with self-attention generative adversarial networks," in *2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, Shanghai, China, 2019.

[8] Z. Cao, G. Hidalgo, T. Simon, S. E. Wei, and Y. Sheikh, "OpenPose: realtime multi-person 2D pose estimation using part affinity fields," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 43, no. 1, pp. 172–186, 2021.

[9] S. Kreiss, L. Bertoni, and A. Alahi, "Pifpaf: composite fields for human pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, 2019.

[10] C. Guo, C. Guo, B. Fan, Q. Zhang, S. Xiang, and C. Pan, "Augfpn: improving multi-scale feature learning for object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA, 2020.

[11] S. Qiao, L.-C. Chen, and A. Yuille, "DetectoRS: detecting objects with recursive feature pyramid and switchable atrous convolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Nashville, TN, USA, 2021.

[12] C.-J. Chou, J.-T. Chien, and H.-T. Chen, "Self adversarial training for human pose estimation," in *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Honolulu, HI, USA, 2018.

[13] M. Andriluka, S. Roth, and B. Schiele, "Pictorial structures revisited: people detection and articulated pose estimation," in *2009 IEEE conference on computer vision and pattern recognition*, Miami, FL, USA, 2009.

[14] S. Johnson and M. Everingham, "Learning effective human pose estimation from inaccurate annotation," in *CVPR 2011*, Colorado Springs, CO, USA, 2011.

[15] Y. Yang and D. Ramanan, "Articulated pose estimation with flexible mixtures-of-parts," in *CVPR 2011*, Colorado Springs, CO, USA, 2011.

[16] L. Pishchulin, L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele, "Poselet conditioned pictorial structures," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, Portland, OR, USA, 2013.

[17] B. Sapp and B. Taskar, "Modec: multimodal decomposable models for human pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Portland, OR, USA, 2013.

[18] A. Toshev and C. Szegedy, "Deeppose: human pose estimation via deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, Columbus, OH, USA, 2014.

[19] S.-E. Wei, S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016.

[20] K. Sun, K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, 2019.

[21] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *European Conference on Computer Vision*, Springer, 2016.

[22] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang, "Multi-context attention for human pose estimation," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, July 2017.

[23] D. Yu, K. Su, J. Sun, and C. Wang, "Multi-person pose estimation for pose tracking with enhanced cascaded pyramid network," *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018Springer, 2018.

[24] B. Cheng, B. Xiao, J. Wang, H. Shi, T. S. Huang, and L. Zhang, "Higherhrnet: scale-aware representation learning for bottom-up human pose estimation," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, June 2020.

[25] Z. Wang, B. Yin, Q. Peng et al., "Rethinking on multi-stage networks for human pose estimation," 2019, https://arxiv.org/abs/1901.00148.

[26] B. M. H. Romeny, *Geometry-Driven Diffusion in Computer Vision*, vol. 1, Springer Science & Business Media, 2013.

[27] S.-W. Kim, S.-W. Kim, H.-K. Kook, J.-Y. Sun, M.-C. Kang, and S.-J. Ko, "Parallel feature pyramid network for object detection," in *Proceedings of the European Conference on Computer Vision (ECCV)*, Munich, Germany, 2018.

[28] C.-F. Chen, Q. Fan, N. Mallinar, T. Sercu, and R. Feris, "Big-little net: an efficient multi-scale feature representation for visual and speech recognition," 2018, https://arxiv.org/1807.03848.

[29] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, https://arxiv.org/abs/1706.05587.

[30] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[31] G. K. Dziugaite, D. M. Roy, and Z. Ghahramani, "Training generative neural networks via maximum mean discrepancy optimization," 2015, https://arxiv.org/abs/1505.03906.

[32] J.-Y. Zhu, P. Krähenbühl, E. Shechtman, and A. A. Efros, "Generative visual manipulation on the natural image manifold," *European Conference on Computer Vision*, 2016Springer, 2016.

[33] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, and L. Van Gool, "Pose guided person image generation," 2017, https://arxiv.org/abs/1705.09368.

[34] X. Peng, Z. Tang, F. Yang, R. S. Feris, and D. Metaxas, "Jointly optimize data augmentation and network training: adversarial data augmentation in human pose estimation," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, June 2018.

[35] I. Goodfellow, I. Goodfellow, J. Pouget-Abadie et al., "Generative adversarial nets," *Advances in Neural Information Processing Systems*, vol. 27, 2014.

[36] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2015, https://arxiv.org/abs/1511.06434.

[37] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," *International conference on machine learning*, 2017PMLR, 2017.

[38] I. Gulrajani, I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of Wasserstein GANs," *NIPS*, 2017.

[39] D. Berthelot, T. Schumm, and L. Metz, "BEGAN: boundary equilibrium generative adversarial networks," 2017, https://arxiv.org/abs/1703.10717.

[40] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," 2017, https://arxiv.org/abs/1710.10196.

[41] H. Zhang, H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," *International conference on machine learning*, 2019PMLR, 2019.

[42] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, https://arxiv.org/abs/1411.1784.

[43] J. Xu, M. Zhang, E. A. Turk, P. E. Grant, P. Golland, and E. Adalsteinsson, "3D fetal pose estimation with adaptive variance and conditional generative adversarial network," in *Medical Ultrasound, and Preterm, Perinatal and Paediatric Image Analysis*, pp. 201–210, Springer, 2020.

[44] Y. Ji, H. Zhang, and Q. M. Jonathan Wu, "Saliency detection via conditional adversarial image-to-image network," *Neurocomputing*, vol. 316, pp. 357–368, 2018.

[45] M. Rezaei, H. Yang, K. Harmuth, and C. Meinel, "Conditional generative adversarial refinement networks for unbalanced medical image semantic segmentation," in *2019 IEEE Winter*

Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, January 2019.

[46] Z. Yuan, H. Li, J. Liu, and J. Luo, "Multiview scene image inpainting based on conditional generative adversarial networks," *IEEE Transactions on Intelligent Vehicles*, vol. 5, no. 2, pp. 314–323, 2019.

[47] P. Mishra and I. Herrmann, "GAN meets chemometrics: segmenting spectral images with pixel2pixel image translation with conditional generative adversarial networks," *Chemometrics and Intelligent Laboratory Systems*, vol. 215, article 104362, 2021.

[48] F. Yu, V. Koltun, and T. Funkhouser, "Dilated residual networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, July 2017.

[49] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley, "Least squares generative adversarial networks," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, October 2017.

[50] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved training of Wasserstein GANs," 2017, https://arxiv.org/abs/1704.00028.

[51] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2D human pose estimation: new benchmark and state of the art analysis," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, June 2014.

[52] T.-Y. Lin, M. Maire, S. Belongie et al., "Microsoft COCO: common objects in context," *European Conference on Computer Vision*, 2014Springer, 2014.

[53] Z. Chen, X. Qin, C. Yang, and L. Zhang, "Composite localization for human pose estimation," 2021, https://arxiv.org/abs/2105.07245.

[54] F. Zhang, X. Zhu, H. Dai, M. Ye, and C. Zhu, "Distribution-aware coordinate representation for human pose estimation," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, June 2020.

[55] A. Bulat, J. Kossaifi, G. Tzimiropoulos, and M. Pantic, "Toward fast and accurate human pose estimation via soft-gated skip connections," in *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, Buenos Aires, Argentina, November 2020.

[56] W. McNally, K. Vats, A. Wong, and J. McPhee, "EvoPose2D: pushing the boundaries of 2D human pose estimation using neuroevolution," 2020, https://arxiv.org/abs/2011.08446.

[57] B. Xiao, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," *Proceedings of the European conference on computer vision (ECCV)*, Springer, 2018.

[58] J. Carreira, P. Agrawal, K. Fragkiadaki, and J. Malik, "Human pose estimation with iterative error feedback," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, June 2016.

[59] P. Hu and D. Ramanan, "Bottom-up and top-down reasoning with hierarchical rectified Gaussians," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, June 2016.

[60] W. Yang, S. Li, W. Ouyang, H. Li, and X. Wang, "Learning feature pyramids for human pose estimation," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, October 2017.

[61] L. Ke, M.-C. Chang, H. Qi, and S. Lyu, "Multi-scale structure-aware network for human pose estimation," *Proceedings of the european conference on computer vision (ECCV)*, Springer, 2018.

[62] W. Tang, P. Yu, and Y. Wu, "Deeply learned compositional models for human pose estimation," *Proceedings of the European conference on computer vision (ECCV)*, Springer, 2018.

[63] T. Sekii, "Pose proposal networks," *Proceedings of the European Conference on Computer Vision (ECCV)*, Springer, 2018.

[64] F. Zhang, X. Zhu, and M. Ye, "Fast human pose estimation," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, June 2019.

[65] M. Fieraru, A. Khoreva, L. Pishchulin, and B. Schiele, "Learning to refine human pose estimation," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Salt Lake City, UT, USA, June 2018.

[66] U. Iqbal and J. Gall, "Multi-person pose estimation with local joint-to-person associations," *European Conference on Computer Vision*, 2016Springer, 2016.

[67] E. Levinkov, J. Uhrig, S. Tang et al., "Joint graph decomposition & node labeling: problem, algorithms, applications," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, July 2017.

[68] E. Insafutdinov, M. Andriluka, L. Pishchulin et al., "ArtTrack: articulated multi-person tracking in the wild," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, July 2017.

[69] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun, "Cascaded pyramid network for multi-person pose estimation," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, June 2018.

[70] G. Papandreou, T. Zhu, N. Kanazawa et al., "Towards accurate multi-person pose estimation in the wild," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, July 2017.