A COMPREHENSIVE FRAMEWORK FOR OCCLUDED HUMAN POSE ESTIMATION

Linhao Xu[1], Lin Zhao[1(✉)], Xinxin Sun[1], Di Wang[2], Guangyu Li[1], Kedong Yan[1]

1. Nanjing University of Science and Technology, Nanjing, China
2. Xidian University, Xi'an, China

ABSTRACT

Occlusion presents a significant challenge in human pose estimation. The challenges posed by occlusion can be attributed to the following factors: 1) Data: The collection and annotation of occluded human pose samples are relatively challenging. 2) Feature: Occlusion can cause feature confusion due to the high similarity between the target person and interfering individuals. 3) Inference: Robust inference becomes challenging due to the loss of complete body structural information. The existing methods designed for occluded human pose estimation usually focus on addressing only one of these factors. In this paper, we propose a comprehensive framework DAG (Data, Attention, Graph) to address the performance degradation caused by occlusion. Specifically, we introduce the mask joints with instance paste data augmentation technique to simulate occlusion scenarios. Additionally, an Adaptive Discriminative Attention Module (ADAM) is proposed to effectively enhance the features of target individuals. Furthermore, we present the Feature-Guided Multi-Hop GCN (FGMP-GCN) to fully explore the prior knowledge of body structure and improve pose estimation results. Through extensive experiments conducted on three benchmark datasets for occluded human pose estimation, we demonstrate that the proposed method outperforms existing methods. Code and data will be publicly available.

Index Terms— Human Pose Estimation, GCN, Occlusion Scenes Analysis

1. INTRODUCTION

Human pose estimation (HPE) has been a prominent area of research in computer vision, with the primary goal of accurately localizing annotated keypoints of human body, such as wrists and eyes, etc. This fundamental task serves as a basis for numerous downstream applications, including human action recognition[1],human-computer interaction [2] and pedestrian re-identification[3], etc.

Thanks to the powerful nonlinear mapping capabilities offered by neural networks, HPE has experienced notable advancements in recent years. However, existing methods such

(a) Keypoints Swap   (b) Keypoints Miss

Fig. 1: Challenges existed in occluded pose estimation. (a)Keypoints Swap: Due to the interference of the non-target person, some keypoints are predicted in the wrong positions. (b)Kypoints Miss: The large area of occlusion leads the contextual and structural information loss. Left is ground truth results and right is detection results.

as PPNet[4], TokenPose[5], ViTPose[6] and PoseTrans[7] all encounter challenges in addressing occlusion.

The primary difficulties of occluded human pose estimation mainly include the invisibility of occluded body parts and the strong interference caused by non-target keypoints. To be specific, there are three main factors: 1) Lack of occluded samples. Mainstream datasets of HPE lack occluded samples, which limits the ability of models to learn robust representations when facing occlusion. 2) Feature confusion. As shown in Fig.1(a). Occlusion can lead to feature confusion because of the high similarity between the target and interfering persons, resulting in perplexity between the target and interfering keypoints. 3) Inference with uncompleted context information. As shown in Fig.1(b). The large area occlusion leads to the loss of contextual and structural information, thus making the model unable to obtain enough contextual information from adjacent keypoints to infer the exact location, which leads to keypoints missing or abnormal postures.

In response to the aforementioned challenges, several methods have been proposed. For instance, ASDA[8] and MSR-Net[9] try to generate occlusion samples, thereby mitigating the impact of limited occlusion samples in the dataset. STIP [10] attempts to address occlusion by enhancing keypoints' semantic information. Besides, methods like OAS-Net[11] focus on optimizing the processing of features to alleviate confusion between the target person and interfering individuals. Furthermore, to overcome the incomplete human body structure information caused by occlusion, approaches such as OPEC-Net[12] leverages the graph convolutional net-
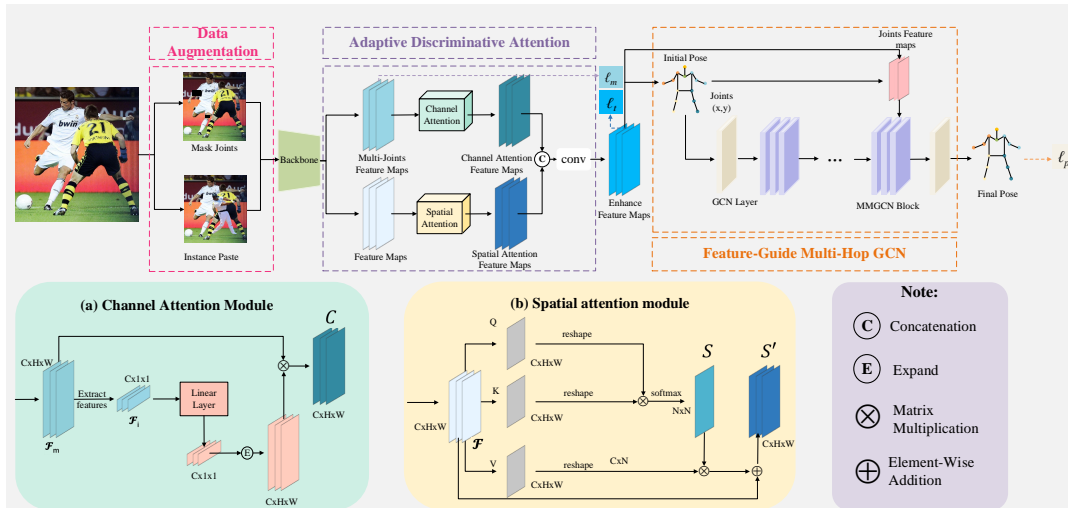
Fig. 2: The framework of our proposed method DAG. The input image undergoes data augmentation, then it is fed into the backbone for feature extraction. Subsequently, the features are input into the adaptive discriminative attention module for feature enhancement, the enhanced features are generated for initial pose generation. The initial pose is then sent to the feature-guided multi-hop GCN for pose refinement and correction, producing the final pose.

work (GCN) to infer occluded keypoints. However, these methods only address one aspect of the challenges, thus their performance is still not ideal when facing occlusion.

In this paper, we present a comprehensive framework DAG (Data, Attention, Graph) to tackle the issue of occlusion in human pose estimation. First, we propose a novel data augmentation method called mask joints with instance paste. Unlike previous methods that focus on simulating either object occlusion or human occlusion alone, our method is compatible with both scenarios, allowing us to better simulate real-world occlusion situations. Second, in order to distinguish confused features caused by occlusion, an adaptive Discriminative Attention Module (ADAM) is introduced to differentiate target person and interfering individuals and enhance target features. Third, to compensate for incomplete body structure information, the Feature-Guided Multi-Hop GCN (FGMH-GCN) is introduced. FGMH-GCN can fully explore the prior knowledge of body structure and leverage useful information in the features map to compensate for the information loss in the initial pose estimation. Our method encompasses data augmentation, feature processing, and result refinement, providing a holistic approach to address occlusion-related difficulties.

2. METHOD

2.1. Overview

As shown in Fig.2, our framework DAG is composed of three main components: Mask Joints with Instance Paste, Adaptive Discriminative Attention Module(ADAM), and Feature-Guided Multi-Hop GCN (FGMP-GCN).

2.2. Mask Joints with Instance Paste

We randomly mask partial keypoints to simulate object occlusion. On the other hand, person occlusion is simulated using instance paste, where occluding instances are inserted into the image.

Mask Joints. In scenarios where a person is occluded by an object, certain parts of the human body become invisible, so we use the joints mask method to simulate occlusion by objects, as shown in Fig.2 . The process involves randomly selecting human and masking keypoints with rectangles of various sizes. This simulation enhances the robustness of the model by exposing it to occlusion challenges during training.

Instance Paste. Inspired by ASDA[8] and in order to mitigate the adverse effects of person-to-person occlusion, we propose instance-paste to enhance the network's robustness. We begin by segmenting all human body images using the human parsing method[13]. These segmented human bodies form our human body instances pool. Human instances are randomly selected from the instances pool with random rotation and scaling to create occlusion data. Finally, the selected human instances are pasted into the image at random positions. This process generates diverse person patterns that mimic real-world scenarios.

2.3. Adaptive Discriminative Attention

To address the challenge of feature confusion arising from occlusion, we introduce the ADAM which comprises two components: channel attention and spatial attention.

Channel Attention. The channel attention leverages the central features of the human body to enhance the corresponding features, thereby facilitating the discrimination between different human instances at the channel level.

Given the multi-joints features $\mathcal{F}_m \in \mathbb{R}^{C \times H \times W}$ output from the backbone. H and W are the height and width of the features map, respectively. And the instance feature $\mathcal{F}_i \in \mathbb{R}^{C \times 1 \times 1}$ is extracted from the center of the human body in $\mathcal{F}_m$. Then F_i is fed into a linear layer to generate new features with the shape as $\mathcal{F}_i$. After that, an expansion operation is used to make the new features resemble $\mathcal{F}_m$. So the channel attention map $C \in \mathbb{R}^{C \times H \times W}$ can be defined as:

$$C = \mathcal{F}_m \cdot \Psi(\mathbf{L}(\mathcal{F}_i)), \tag{1}$$

$\mathbf{L}$ is a linear transformation, and Ψ is the expansion operation.

Spatial Attention. Spatial attention is leveraged to suppress background features or object features that may occlude the human body.

Given the feature maps $\mathcal{F} \in \mathbb{R}^{C \times H \times W}$, we first use convolution layers to generate three new feature maps Q, K, V, where $\{Q, K, V\} \in \mathbb{R}^{C \times H \times W}$. Next, $\{Q, K\}$ are reshaped to $\mathbb{R}^{C \times N}$, where $N = H \times W$ represents the number of pixels. Following this, the transpose of K is multiplied by Q to obtain the spatial attention map $S \in \mathbb{R}^{N \times N}$. After reshaping V to $\mathbb{R}^{C \times N}$, a matrix multiplication is performed between V and the transpose of S. The resulting matrix is then reshaped back to $\mathbb{R}^{C \times H \times W}$, yielding the final result $S' \in \mathbb{R}^{C \times H \times W}$:

$$S' = \Phi(\mathcal{F})_V \cdot softmax((\Phi(\mathcal{F})_K^T \odot \Phi(\mathcal{F})_Q))^T. \tag{2}$$

where Φ is a convolution operation.

2.4. Feature-Guided Multi-Hop Graph convolutional Network

GCN fomulation We construct a graph $G = (\mathcal{V}, \mathcal{E})$ to formulate the human pose with N joints. Here, $\mathcal{V}$ represents each keypoint of the human body, and $\mathcal{E}$ represents the connections between two keypoints in the body. The collection of features of all nodes can be written as a matrix $M \in \mathbb{R}^{D \times N}$. The updated feature matrix can be written as:

$$M' = \sigma(\sum_{k=1}^{K} w_k \cdot (W^{(0)}M + W^{(1)}M\hat{A}_k)). \tag{3}$$

Where $w_k \in \mathbb{R}^{D \times N}$ is a learnable modulation matrix used to model the relationship between features at each hop distance, $W^{(0)}$ and $W^{(1)}$ are the weight matrices corresponding to the self and neighbor transformations, respectively. $\hat{A}$ is the symmetrically normalized version without self-connections. D is the dimension of features.

Feature-Guided. The MMGCN block [14] is employed to capture valuable information present in the features map that may have been lost in the initial pose. Next, the grid sampling is utilized to extract joint features from their respective feature map locations (x, y). Then the joint features are weighted into the original features map.

Multi-Hop. Traditional graph convolutional networks (GCNs) often only consider information between adjacent nodes. Inspired by MMGCN [14], we introduce the mechanism of multi-hop. The relationship in the adjacency matrix

of each hop is low correlation except for the distance within k hops of the intermediate nodes. Therefore, it can provide a flexible modeling structure for learning the long-term relationships between human joints. FGMH-GCN not only makes up for the missing human structure information by using the prior knowledge of body structure but also suppresses the generation of abnormal poses.

2.5. Loss Function

The loss function consists of three parts 1) multi-joints loss ℓ_m; 2) target-joints loss ℓ_t; 3) GCN-pose loss ℓ_p. For the keypoint heatmap, we use mean square error (MSE) as our loss function. For the GCN pose loss, we use L_1 loss. So the overall loss function can be written as:

$$\mathcal{L} = \ell_m + \ell_t + \lambda \ell_p. \tag{4}$$

Where λ is a hyperparameter for balancing the three losses.

3. EXPERIMENTS

3.1. Datasets

We evaluate our model on three datasets: MSCOCO-RE, CrowdPose [19] and OCHuman [20].

COCO-RE the train2017 dataset contains 57K images with over 150K person instances. val2017 set and test-dev set contain 5K images and 20K images, respectively. COCO-RE is our re-labeling of the val2017 set, which adds the annotations of the occluded joints. The visible flag is denoted as $v = 3$: labeled and occluded.

CrowdPose [19] contains 20K images and 80K persons labeled with 14 keypoints. We use the trainval set for training and the test set for evaluation.

OCHuman [20] contains 4731 images and 8110 person instances. It consists of 2,500 images in the val set and 2,231 images in the test set.

3.2. Implementation Details

All experiments are conducted using PyTorch on two RTX 3090 GPUs. The HRNet is used as the baseline model and is initialized with weights pre-trained on the ImageNet classification task.

Training. We train all models using the HRNet framework, wherein the human bounding box is extended in a fixed aspect ratio of height:width=4:3, and the region is cropped from the original image which is then resized to a fixed size of 256x192 or 384x288. We follow the HRNet for other related settings.

Testing. We follow the top-down workflow for human pose estimation. In the case of the MSCOCO dataset, we use the detection results of previous works[16] to ensure a fair comparison. As for the CrowdPose dataset, we utilize ResNet101-FPN[21] as the human detector to detect individuals. The heatmap post-processing is the same as the HRNet.

method	backbone	input size	val2017 AP	test-dev AP	RE AP	OCHuman AP	CrowdPose AP
SBL [15]	ResNet-50	256x192	70.4	70.0	67.1	55.8	68.4
OAS-Net [11]	HRNet-W32	256x192	75.0	-	-	-	-
ASDA [8]	HRNet-W32	256x192	75.2	-	-	-	-
OPEC-Net [12]	HRNet-W32	256x192	-	73.9	-	-	-
HRNet [16]	HRNet-W32	256x192	<u>74.4</u>	<u>73.5</u>	<u>71.0</u>	63.0	<u>71.7</u>
HRNet [16]	HRNet-W48	384x288	<u>76.3</u>	<u>75.5</u>	<u>73.0</u>	64.8	<u>73.9</u>
STIP [10]	HRNet-W48	384x288	76.8	-	-	64.0	-
PoseTrans [7]	HRNet-W48	384x288	76.8	75.7	-	-	-
SimCC [17]	HRNet-W32	256x192	75.3	74.3	71.8	<u>62.3</u>	66.7
SimCC [17]	HRNet-W48	384x288	76.9	76.0	73.5	<u>66.2</u>	-
Poseur [18]	ResNet-50	256x192	<u>74.2</u>	<u>72.8</u>	<u>70.6</u>	<u>58.0</u>	<u>70.9</u>
DAG	HRNet-W32	256x192	**75.4**↑$_{1.0}$	**74.3**↑$_{0.8}$	**72.0**↑$_{1.0}$	**64.5**↑$_{2.2}$	**72.7**↑$_{1.0}$
DAG	HRNet-W48	384x288	**77.0**↑$_{0.7}$	**76.0**↑$_{0.5}$	**73.7**↑$_{0.7}$	**66.9**↑$_{0.7}$	**74.2**↑$_{0.3}$
Poseur + DAG	ResNet-50	256x192	**74.5**↑$_{0.3}$	**73.1**↑$_{0.3}$	**71.1**↑$_{0.5}$	**58.9**↑$_{0.9}$	**71.5**↑$_{0.6}$

Table 1: Comparison with Other methods on MSCOCO val, test-dev, RE, OCHuman, CrowdPose. The <u>underlined</u> highlights the compared results, the results of our method are highlighted in **bold**.

3.3. Comparison with Other Methods

Quantitative results of our method DAG and other methods on MSCOCO, OCHuman and CrowdPose are listed in Tab.1. Compared to HRNet, DAG achieves 0.7% and 0.5% gains on val and test-dev set with HRNet-W48 and input size 384x288. The performance of traditional methods, such as SBL and HRNet, decreases when testing on MSCOCO-RE. In contrast, our DAG achieves a performance of 73.7 AP, which is an improvement of 0.7% mAP compared to HRNet.

On OCHuman, We achieved 2.1% mAP improvement compared to the baseline HRNet-W48 model with an input size of 384x288. Additionally, we achieve 2.2% mAP improvement compared to the Simcc HRNet-W32 model with an input size of 256x192.

On CrowdPose, our proposed DAG has demonstrated a significant improvement of 1.0% mAP compared to the baseline, indicating its effectiveness and generalizability across different scenarios.

Moreover, when combined with our DAG, the Poseur[18] exhibited notable improvements across multiple datasets, especially in the occluded datasets. This provides strong evidence of DAG's effectiveness, which also shows DAG can be easily integrated into other frameworks, making it a versatile solution for improving human pose estimation performance in occlusion scenarios.

3.4. Ablation Study

Ablation Studies on MSCOCO. The results on MSCOCO are summarized in Tab.2. HRNet is used as the baseline with an input size of 256x192. Our proposed method improves the mAP by 1% compared to the baseline. Specifically, the data augmentation technique effectively simulates real-world occlusion scenarios and improves the mAP by 0.3%. Moreover, ADAM distinguishes the target from the interference

MSCOCO val2017 datsets				
HRNet-W32	DA	ADAM	FGMH-GCN	AP
✓				74.4
✓	✓			74.7 (0.3 ↑)
✓		✓		74.8 (0.4 ↑)
✓			✓	75.1 (0.7 ↑)
✓	✓	✓	✓	75.4 (1.0 ↑)

Table 2: Ablation study. Investigating the effects of the proposed module. DA means adding the proposed data augmentation module.

and further enhances the performance by 0.4% mAP. Furthermore, FGMH-GCN improves the mAP by 0.7%, effectively capturing the relationship between neighboring keypoints. The improvement induced by the proposed techniques clearly demonstrates the effectiveness of the proposed method.

4. CONCLUSION

In this paper, we propose a comprehensive framework DAG for occluded pose estimation. In particular, we design a novel data augmentation method mask joints with instance paste to generate challenging occluded samples. Additionally, an Adaptive Discriminative Attention Module is introduced to distinguish the confusion features. Furthermore, we incorporate a Feature-Guided Multi-Hop GCN that leverages the human body structure to strengthen the body structure constraints during joint inference. Extensive experiments on three benchmarks demonstrate the effectiveness and generalizability of our method.

5. ACKNOWLEDGMENTS

6. REFERENCES

[1] Bingbing Ni, Teng Li, and Xiaokang Yang, "Learning semantic-aligned action representation," *IEEE transactions on neural networks and learning systems*, vol. 29, no. 8, pp. 3715–3725, 2017. 1

[2] Osama Mazhar, Sofiane Ramdani, and Benjamin Navarro, "Towards real-time physical human-robot interaction using skeleton information and hand gestures," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018, pp. 1–6. 1

[3] Yulin Li, Jianfeng He, and Zhang, "Diverse part discovery: Occluded person re-identification with part-aware transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2898–2907. 1

[4] Lin Zhao, Nannan Wang, Chen Gong, Jian Yang, and Xinbo Gao, "Estimating human pose efficiently by parallel pyramid networks," *IEEE Transactions on Image Processing*, vol. 30, pp. 6785–6800, 2021. 1

[5] Yanjie Li, Shoukui Zhang, and Zhicheng Wang, "Tokenpose: Learning keypoint tokens for human pose estimation," in *Proceedings of the IEEE International conference on computer vision*, 2021, pp. 11313–11322. 1

[6] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao, "ViTPose: Simple vision transformer baselines for human pose estimation," in *Advances in Neural Information Processing Systems*, 2022. 1

[7] Wentao Jiang, Sheng Jin, and Wentao Liu, "Posetrans: A simple yet effective pose transformation augmentation for human pose estimation," in *Proceedings of the European conference on computer vision (ECCV)*, 2022, pp. 643–659. 1, 1

[8] Yanrui Bin, Xuan Cao, and Xinya Chen, "Adversarial semantic data augmentation for human pose estimation," in *Proceedings of the European conference on computer vision (ECCV)*, 2020, pp. 606–622. 1, 2.2, 1

[9] Lipeng Ke, Ming-Ching Chang, and Honggang Qi, "Multi-scale structure-aware network for human pose estimation," in *Proceedings of the european conference on computer vision (ECCV)*, 2018, pp. 713–728. 1

[10] Xuanhan Wang, Lianli Gao, and Yan Dai, "Semantic-aware transfer with instance-adaptive parsing for crowded scenes pose estimation," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 686–694. 1, 1

[11] Lu Zhou, Yingying Chen, and Yunze Gao, "Occlusion-aware siamese network for human pose estimation," in *Proceedings of the European conference on computer vision (ECCV)*, 2020, pp. 396–412. 1, 1

[12] Lingteng Qiu, Xuanye Zhang, and Yanran Li, "Peeking into occluded joints: A novel framework for crowd pose estimation," in *Proceedings of the European conference on computer vision (ECCV)*, 2020, pp. 488–504. 1, 1

[13] Tao Ruan, Ting Liu, and Zilong Huang, "Devil in the details: Towards accurate single and multiple human parsing," in *Proceedings of the AAAI conference on artificial intelligence*, 2019, vol. 33, pp. 4814–4821. 2.2

[14] Jae Yung Lee and I Gil Kim, "Multi-hop modulated graph convolutional networks for 3d human pose estimation," in *Proceedings of the British Machine Vision Conference*, 2022, pp. 1–13. 2.4

[15] Bin Xiao, Haiping Wu, and Yichen Wei, "Simple baselines for human pose estimation and tracking," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 466–481. 1

[16] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang, "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5693–5703. 3.2, 1

[17] Yanjie Li, Sen Yang, and Peidong Liu, "Simcc: A simple coordinate classification perspective for human pose estimation," in *Proceedings of the European conference on computer vision (ECCV)*, 2022, pp. 89–106. 1

[18] Weian Mao, Yongtao Ge, and Chunhua Shen, "Poseur: Direct human pose regression with transformers," in *Proceedings of the European conference on computer vision (ECCV)*, 2022, pp. 72–88. 1, 3.3

[19] Jiefeng Li, Can Wang, and Hao and Zhu, "Crowdpose: Efficient crowded scenes pose estimation and a new benchmark," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 10863–10872. 3.1

[20] Song-Hai Zhang, Ruilong Li, and Xin Dong, "Pose2seg: Detection free human instance segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 889–898. 3.1

[21] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, and Kaiming He, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125. 3.2