ViPose: Keypoint Visibility-based Human Pose Estimation

Xingyuan Ye[1], Dan Lin[2, *], Kim-Hui Yap[1]

[1]School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore
[2]College of Computer Science and Technology, Harbin Engineering University, Harbin, Heilongjiang, China
Email: XYE006@e.ntu.edu.sg, danlin@hrbeu.edu.cn, EKHYap@ntu.edu.sg
*Corresponding author: danlin@hrbeu.edu.cn

***Abstract*—Human Pose Estimation (HPE) aims to predict the positional coordinates of body keypoints in images. While significant progress has been made in HPE, certain challenges persist. For example, the potential for robust occlusion can result in considerable confidence in predicting false-positive keypoints. This can cause problems in applications where high detection accuracy is required, such as in a traffic environment. Most existing methods focus on predicting the positional coordinates of each human joint, but they fail to consider the keypoint occlusion problem and detect the visibility of keypoints in the image. In this paper, we propose a visibility-guided human pose estimation model, named ViPose. We extract the visibility information of keypoints to detect whether a specific keypoint of the subject in the image is obscured by other objects, visible or not captured. ViPose is composed of two modules: HPE keypoint coordinate predictor module, and keypoint visibility detector module. To verify the effectiveness of ViPose, we conduct experiments on both the general MSCOCO dataset and the diver monitoring DriPE dataset. The experimental results show that ViPose can achieve good performance with fewer parameters.**

***Index Terms*—Human Pose Estimation; Keypoint Detection; Car Cabin Monitoring;**

I. Introduction

Human Pose Estimation (HPE) is a task that focuses on the localization and identification of specific body keypoints for the given images [1]. These keypoints typically include significant anatomical body joints (such as shoulders, elbows, hips, and ankles), as well as facial markers (such as eyes, ears, and nose). In some real-world applications, fine-grained keypoints on the feet, hands, or face may also be utilized [2]. Within the car cabin environment, HPE has become a fundamental task which can provide the primary understanding for the follow-up tasks, such as driver action recognition [3], human-object interaction [4], and so on. In addition, HPE has been widely applied to different scenarios, including motion analysis [5], activity analysis [6], and augmented reality [7].

Recently, deep convolutional neural networks (CNNs) have demonstrated remarkable efficacy in the domain of single-person and multi-person HPE. Li et. al. proposed a graphical model to first build spatial interactions as graphs for 3D HPE [8]. In addition, Khan et al. proposes a human gait recognition framework using deep learning and Bayesian optimization, achieving high accuracy by enhancing video frames and employing a novel feature fusion and selection approach [9]. Both single-person and multi-person existing HPE methods mainly focus on improving the performance of keypoint coordinate prediction directly.

However, a significant challenge faced by most previous HPE methods is their inability to handle occlusions of keypoints. Despite recent advancements in the HPE task, most of the HPE datasets primarily comprise images with minimal occlusion. Strong occlusion may cause the model to predict keypoints with high confidence, which may not exist in the image. Such erroneous predictions of keypoints can pose significant challenges for applications where precise estimation is essential, such as the follow-up human action recognition task and the driver's pose analysis.

One major gap is the limited consideration of keypoint visibility as an auxiliary factor in pose estimation. While some studies have touched upon the importance of keypoint visibility, they often do not fully integrate this aspect into their estimation models. Existing methods although effective in certain scenarios, lack robustness in cases where keypoints are partially or entirely occluded. This oversight can lead to significant inaccuracies in pose estimation, particularly in complex real-world environments.

To solve the challenge above, we propose a keypoint visibility-based human pose estimation model, named ViPose. The main contribution of this paper is listed as follow:

- We combine human pose estimation and keypoint visibility detector to enhance the accuracy of HPE predictions.
- ViPose detects and identifies the visibility of keypoints of the subject with three labels: fully visible, not visible, and not labelled. In parallel, ViPose introduces a transformer to detect keypoint coordinates.
- To verify the effectiveness of ViPose, we conduct experiments from various angles using two datasets: the general MSCOCO dataset, and the diver monitoring DriPE dataset.
- The experimental results show that the ViPose model outperforms the previous models in both datasets. In addition, ViPose can achieve good performance with less parameters.

979-8-3503-6463-7/24/$31.00 ©2024 IEEE

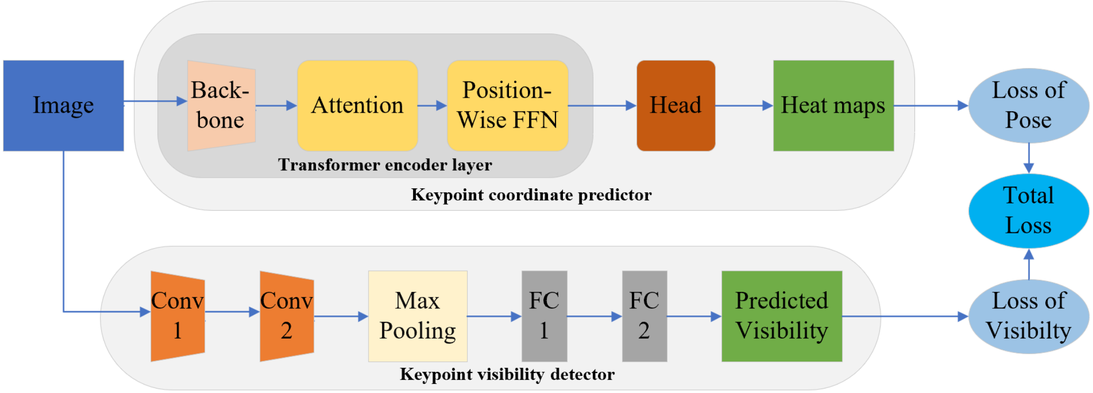

Fig. 1: The framework of the proposed ViPose model. The keypoint coordinate predictor introduces a Transformer encoder for human pose estimation, and the keypoint visibility predictor utilizes the CNN to detect the visibility of each keypoint. ‘Conv1’ and ‘Conv2’ represent two CNN layers. ‘FC1’ and ‘FC2’ denote two fully connected layers.

II. Related work

A. Human Pose Estimation

According to the individuals of the image, the HPE task can be divided into two categories: single-person HPE task and multi-person HPE task. The single-person HPE concentrates on detecting human poses in images with one individual, commonly in top-down approach such as tree-structured graphical models and global-local-graphical models [10].

In contrast, multi-person HPE task detects multiple coordinates for several individuals within the image. The bottom-up approach is widely employed in multi-person HPE [11]. This strategy first detects the individuals and then estimates the pose of each individual independently in each detected region. A significant advantage is that the computation consumption for each image is only dependent on the image’s resolution and regardless of the number of individuals present in the given image. BEH⊙LDER combines real-time location systems and deep learning-based multi-person human pose estimation to monitor and improve surgical workflows in hybrid operating suites [12]. CGARPN then enhances human behavior understanding in complex scenarios by integrating global association features and adaptive routing for improved accuracy and efficiency [13]. Recently, a human gait recognition framework achievs high accuracy by enhancing video frames and employing a novel feature fusion and selection approach [9]. However, these strategies suffers from limited performance due to the complex features of different individuals.

B. Keypoint Visibility on HPE

Recent research on HPE mainly concentrated on enhancing the accuracy of keypoint coordinate prediction. However, the analysis of the keypoint visibility is ignored by most studies. State-of-the-art HPE datasets, such as MSCOCO [14], provide visibility labels that indicate the presence of each keypoint.

Some algorithms introduce visibility information to the HPE task. In [15], the authors utilized visibility prediction to introduce a novel evaluation method for multi-person HPE in scenarios where heavy occlusion is present. The predicted visibility of keypoints is quantified as an occlusion score, which is subsequently used to calculate a performance metric specifically designed to highlight how effectively the evaluated networks handle occluded keypoints. Furthermore, the multi-instance HPE network incorporated a Transformer module to evaluate the keypoint visibility, a feature treated as a secondary task within the end-to-end training process [16]. However, the visibility is a binary score and cannot present the keypoints that are not labelled. Golda et. al. proposed OccNet, which can predict the occlusion in the scenario where many people are occluded in the images [17]. Guesdon et. al. added a visibility predictor to the HPE base model [18]. However, this model suffers from sub-optimal performance due to the rough feature extraction for the human pose.

Although keypoint visibility has been investigated in HPE, there are still several challenges. The aforementioned methods have focused on predicting binary visibility and failed to represent the comprehensive visibility labels in existing datasets. These visibility labels include non-visible, visible, and non-labelled. In addition, the previous studies provide limited quantitative results on the efficacy of visibility predictions. And the visibility prediction component is primarily auxiliary in the proposed fixed network. To address these limitations, we present a modified model that enables HPE techniques to utilize both keypoints coordinates and keypoint visibility.

III. Methodology

The detailed framework of the ViPose model is illustrated in Fig. 1. The ViPose model is composed of two modules, the HPE keypoint coordinate predictor module as the base model, and the keypoint visibility detector module. Then in the training phase, these two modules are integrated to achieve improved performance for final pose estimation.

A. Keypoint coordinate predictor

As shown in Fig. 1, the keypoint coordinate predictor consists of three parts: the image feature extraction backbone, keypoint extraction transformer, and keypoint prediction head. Specifically, ViPose first utilizes ResNet-S [19] as the backbone to extract the features of input images. Then, a transformer encoder is introduced for human pose feature extraction. Finally, the output of the transformer encoder is augmented with a classification head to estimate human keypoint heatmaps. The detailed structure is introduced in the following.

To reduce the parameter redundancy, the ResNet-S [19], a simplified version of ResNet [20], is employed as the backbone in this paper. ResNet-S requires much less parameters of the original ResNet, making ViPose more adaptable to real-world applications in car cabinenvironments. For the input image $I \in \mathbb{R}^{3\times H\times W}$, ResNet-S generates a 2D spatial feature map of the image $\mathbf{X}_f \in \mathbb{R}^{d\times H\times W}$. Then, the feature map of the image is transformed into a sequence $\mathbf{X} \in \mathbb{R}^{L\times d}$ by the flattening operation, where $L = H \times W$.

Then, we introduce the transformer encoder to learn the keypoint features. To clarify, only the encoder of the transformer is used in ViPose. This is because the task of final heatmaps prediction can be viewed as an encoding task.The flattened sequence $\mathbf{X}$ is subsequently processed through a sequence of N attention layers and feed-forward networks (FFNs). In each attention layer, $\mathbf{X}$ is projected into sequence $\mathbf{Q} \in \mathbb{R}^{L\times d}$. Then, a multi-head attention mechanism is deployed. The computation of attention scores matrix $\mathbf{A} \in \mathbb{R}^{N\times N}$ is as [19]:

$$\mathbf{A} = \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right), \tag{1}$$

where $\mathbf{K} \in \mathbb{R}^{L\times d}$ denotes the keys, and d denotes the dimension of image feature X.

The output of the transformer encoder $\mathbf{E} \in \mathbb{R}^{L\times d}$ is augmented with a head to estimate K different types of human joints heatmaps $P \in \mathbb{R}^{K\times H^*\times W^*}$, where H^* and W^* equals to the input size $\frac{H}{4}$ and $\frac{W}{4}$, respectively. To achieve the output, the matrix $\mathbf{E}$ is first reshaped to $\mathbb{R}^{d\times H\times W}$, aligning it with the spatial dimensions required for convolutional processing. Then, a CNN layer with a kernel size of 1×1 is applied. This specific kernel size effectively projects the dimensionality of the matrix from d to K without altering the spatial dimensions H and W. If the dimensions H and W do not match the desired dimensions H^* and W^*, it is necessary to adjust the size through upsampling techniques before the 1×1 convolution. We employ a supplementary linear interpolation or a 4×4 transformed convolution, depending on the specific requirements for the application and the data.

B. Keypoint visibility detector

Parallel to the keypoint coordinate predictor, we design a visibility detector to predict the existence of the keypoint in the given image. We adhere to the formalism used in the MSCOCO dataset and define visibility ground truth with integer labels: 0 denotes the keypoint is not labelled, 1 denotes labelled but not visible, and 2 denotes fully visible. Thus, each keypoint is assigned one of the three labels.

The detailed structure of the keypoint visibility detector is as follows. For the input images, We first utilize two CNN layers (Conv1 and Conv2 in Fig. 1) to extract the features, followed by a max pooling layer. Then, the network stacks two fully connected layers (FC1 and FC2 in Fig. 1) as the classifier to map the extracted features to the output label. The final output is a 1×17 vector representing the visibility prediction results for 17 keypoints in each image. The prediction results of the visibility detector also obey the three values of ground truth.

C. Loss Function

ViPose is trained in an end-to-end way [18]. The global loss function is the combination of the keypoint coordinate predictor loss L_T and the keypoint visibility detector loss L_V:

$$\boldsymbol{L} = (1-\alpha)\cdot \boldsymbol{L}_T + \alpha \cdot \boldsymbol{L}_V, \tag{2}$$

where the parameter α is trained to balance the ratio between the loss functions L_T and L_V. Both two modules utilize Mean Square Error applied to the predictions and the ground truth. In subsequent experiments, various α values are investigated for optimal performance.

IV. Experiments and results

A. Implementation details

For the keypoint coordinate predictor module, the attention receives $1/8$ down-sampling resolution, the number of heads is 8, and the number of the attention layers in the Transformer encoder is set to 4.

For the visibility detector, two CNN layers are with a kernel size of 3×3. The first layer expands the channel of the image from 3 to 32 and the second layer further expands it to 64. The kernel size of the max pooling layer is 2×2. The fully connected layer finally generates the output with a dimension of 17×1, representing the visibility of each keypoints.

The resolution of the input images is uniformly adjusted to 256×192. The generated heatmaps resolution is with a dimension of 64×48. To determine the optimal parameter α, we select the values from $\{0, 0.1, 0.2, 0.3\}$ and evaluate the performance under these values.

B. Datasets and evaluation metrics

We evaluate the ViPose model on two datasets: MSCOCO dataset [14] and the DriPE dataset [24].

MSCOCO. This dataset is one of the most commonly used datasets for computer vision research. The MSCOCO dataset comprises 200k images captured in natural settings, with a total of 250k person instances in total. For the HPE task, the dataset is divided into train2017, val2017 and test2017 subsets. The train2017 subset consists of 57k images and 150k person instances, while the val2017 set has 5k images. The test2017 subset, used for evaluation, comprises 20k images. Each human instance is annotated with 17 keypoints.

DriPE. Specially designed for vehicle cabin monitoring, this dataset is built from real-world driving scenarios, characterized by complex conditions, such as illumination changes, occluding shadows, and moving foreground [24]. DriPE comprises 10,000 images of drivers captured under authentic driving conditions, with 7,400 images designated for training and 2,600 images equally allocated for validation and testing.

TABLE I: Comparison with state-of-the-art CNN-based models on the MSCOCO dataset.

Method	Parameters	**AP**	**AP**$_{0.5}$	**AP**$_{0.75}$	**AP**$_{M}$	**AP**$_{L}$
SimpleBaseline [21]	68.6M	73.7	91.9	81.1	70.3	80.0
HRNet-W32 [22]	28.5M	74.9	92.5	82.8	71.3	80.9
HRNet-W48 [22]	63.6M	74.2	92.4	82.4	70.9	79.7
DarkPose [23]	63.6M	76.2	92.5	83.6	72.5	82.4
TransPose_R [19]	4.74M	75.1	92.6	82.6	71.9	79.6
ViPose (Ours)	11.2M	75.2	92.5	82.6	72.1	79.9

TABLE II: Comparison with state-of-the-art CNN-based models on the DriPE dataset.

Method	Parameters	**AP**	**AP**$_{0.5}$	**AP**$_{0.75}$	**AR**	**AR**$_{0.5}$	**AR**$_{0.75}$
SBI [24]	71.2M	96.5	99.9	99.9	97.5	99.9	99.9
EfficientNet B0 [25]	55.6M	91.8	99.0	99.0	94.7	99.9	99.6
EfficientNet B6 [25]	95.5M	99.4	99.0	99.0	96.5	99.9	99.6
MSPN 2-stg [26]	104.6M	97.8	99.0	99.0	99.0	99.9	99.9
TransPose_R [19]	4.74M	98.2	100	99.9	98.9	100	99.9
ViPose (Ours)	11.2M	98.4	99.0	99.0	97.5	100	100

The annotation method and data structure employed in the DriPE dataset are identical to those used in MSCOCO.

Following [19], we evaluate the model with the mAP metric [24], as well as average precision (AP), and average recall (AR). AP$_{0.5}$ and AP$_{0.75}$ are the AP with 50% and 75% minimum of Intersection over Union(IoU) respectively. AR$_{0.5}$ and AR$_{0.75}$ are the AR with 50% and 75% minimum of IoU. During the evaluation process, only the data points deemed to be present in the keypoints are included in the results, whereas the points believed to be absent from the image are not taken into account.

C. Model Training Details

We trained our model with Intel® Xeon® W-2295 CPU and a single Nvidia® GeForce® RTX2080Ti GPU, with 11GB VRAM. The batch size is set to 20. The first training stage includes 230 epochs, with an initial learning rate of 1e-4 and a decay factor of 0.25 for each 5 epochs. To avoid overfitting the model, we set the dropout rate to 0.1.

After the initial training phase, we adjust the learning rate to begin at 1e-5, gradually decaying to 1e-6 by the end of the fine-tuning phase.

V. Experimental Result

We evaluate ViPose on both MSCOCO and DriPE datasets, comparing its performance with other HPE models.

A. Comparative experimental results on MSCOCO dataset

Experimental results on MSCOCO are listed in Table I. We compare ViPose to SimpleBaseline [21], HRNet [22], DarkPose [23], and TransPose_R [19]. ViPose outperforms most existing methods in terms of AP in Table I. Although the AP of ViPose is slightly lower than that of DarkPose [23], ViPose achieves this with significantly fewer parameters, approximately 14.7% of those required by DarkPose. The results indicate that ViPose can achieve comparable performance with much less computational complexity.

B. Comparative experimental results on DriPE dataset

To further verify the effectiveness of the proposed ViPose method, we evaluate it on the DriPE dataset. The results are shown in Table II. We compare ViPose with four existing methods: SBI [24], EfficientNet [25], MSPN 2-stg [26], and TransPose_R [19]. ViPose achieves competitive performance on the DriPE dataset, scoring 98.4% AP, slightly lower than EfficientNet B6 [25], but is 8.53 times more efficient. These results highlight ViPose's potential in real-world applications.

C. Ablation study on different α values

We experiment with varying α values from $0, 0.1, 0.2, 0.3$ and compare the results. Table III and Table IV demonstrate the comparative analysis.

TABLE III: Comparison of the performance with different α values on the MSCOCO dataset.

α	**AP**	**AP**$_{0.5}$	**AP**$_{0.75}$	**AR**	**AR**$_{0.5}$	**AR**$_{0.75}$
0	75.1	92.6	82.6	77.8	93.2	84.1
0.1	75.1	92.5	81.6	77.8	93.2	83.8
0.2	75.2	92.5	82.6	77.9	93.4	84.3
0.3	74.4	92.5	81.4	77.1	93.2	83.0

From the results, we can see that the model exhibits the best performance when α is set to 0.2 on the MSCOCO dataset and 0.1 on the DriPE dataset. It is noteworthy that the performance drop is more pronounced as α increases, indicating the importance of balancing pose estimation and keypoint visibility detection. An optimal α value is critical to achieving good performance.

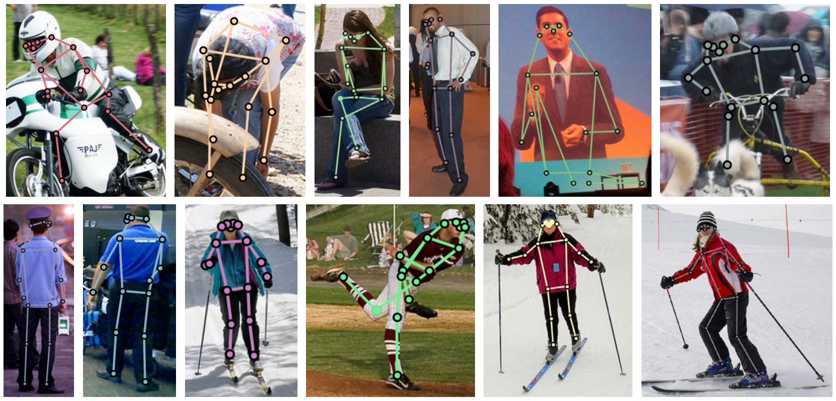

Fig. 2: Some samples of prediction result on the MSCOCO val2017 dataset.

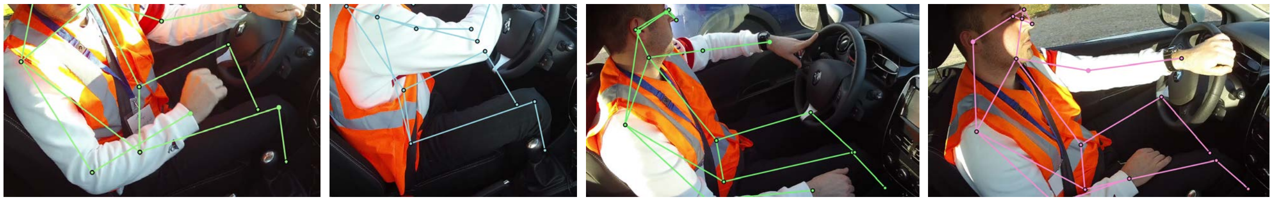

Fig. 3: Some samples of prediction results on the DriPE test dataset.

TABLE IV: Comparison of the performance with different α values on the DriPE dataset.

α	AP	$AP_{0.5}$	$AP_{0.75}$	AR	$AR_{0.5}$	$AR_{0.75}$
0	98.2	100	99.9	98.9	100	99.9
0.1	98.4	100	100	99.0	100	100
0.2	95.9	99.0	99.0	97.4	99.8	99.5
0.3	96.2	99.0	99.0	97.5	99.8	99.5

D. Visualization results and analysis

We utilize COCOAPI to visualize the results predicted on the MSCOCO and the DriPE datasets.

As shown in Fig. 2, the diverse range of human activities and varying conditions in the MSCOCO dataset increases the complexity of pose estimation. Despite this, ViPose successfully estimates human poses, demonstrating its adaptability to diverse and challenging environments.

Fig. 3 shows the results on DriPE dataset on a car cabin environment with a driver in the driving seat. While the pose complexity is lower than that in MSCOCO, DriPE presents challenges with serious occlusions of body parts. Despite these challenges, ViPose achieves surpassing performance, indicating its robustness and suitability for real-world applications.

VI. Conclusion

In this paper, we focus on the occlusion problem of the human pose estimation task and propose a visibility-guided human pose estimation model, named ViPose. Parallel to the HPE keypoint coordinate predictor, the keypoint visibility detector can determine the keypoints of the visible human body, slightly occluded or out of range. By combining the visibility detector and coordinate predictor, ViPose can improve prediction accuracy for the traditional HPE task. The ViPose model is initially trained and evaluated on the MSCOCO dataset, which is recognized as a benchmark dataset for keypoint detection. To further evaluate the effectiveness of ViPose, we evaluate ViPose on driver pose estimation tasks for vehicle cabin monitoring applications. Experimental results on the DriPE dataset illustrate that ViPose can achieve good performance in different applications while requiring fewer parameters.

Acknowledgement

This study is supported under the RIE2020 Industry Alignment Fund Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s).

References

[1] Shradha Dubey and Manish Dixit. A comprehensive survey on human pose estimation approaches. *Multimedia Systems*, 29:167–195, 2022.

[2] Haoming Chen, Runyang Feng, Sifan Wu, Hao Xu, Feng Zhou, and Zhenguang Liu. 2d human pose estimation: A survey. *ArXiv*, abs/2204.07370, 2022.

[3] Dan Lin, Philip Hann Yung Lee, Yiming Li, Ruoyu Wang, Kim-Hui Yap, Bingbing Li, and You Shing Ngim. Multi-modality action recognition based on dual feature shift in vehicle cabin monitoring. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6480–6484, 2024.

[4] Jianjun Gao, Kim-Hui Yap, Kejun Wu, Duc Tri Phan, Kratika Garg, and Boon Siew Han. Contextual human object interaction understanding from pre-trained large

language model. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 13436–13440. IEEE, 2024.

[5] Wen-Li Wei, Jen-Chun Lin, Tyng-Luh Liu, and Hong-Yuan Mark Liao. Capturing humans in motion: Temporal-attentive 3d human pose and shape estimation from monocular video. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13201–13210, 2022.

[6] Liangchen Song, Gang Yu, Junsong Yuan, and Zicheng Liu. Human pose estimation and its application to action recognition: A survey. *Journal of Visual Communication and Image Representation*, 76:103055, 2021.

[7] Mir Suhail Alam, Malik Arman Morshidi, Teddy Surya Gunawan, Rashidah Funke Olanrewaju, and Fatchul Arifin. Pose estimation algorithm for mobile augmented reality based on inertial sensor fusion. *International Journal of Electrical and Computer Engineering (IJECE)*, 2022.

[8] Wenhao Li, Hong Liu, Tianyu Guo, Haoling Tang, and Runwei Ding. Graphmlp: A graph mlp-like architecture for 3d human pose estimation. *ArXiv*, abs/2206.06420, 2022.

[9] Muhammad Attique Khan, Habiba Arshad, Wazir Zada Khan, Majed Alhaisoni, Usman Tariq, Hany S Hussein, Hammam Alshazly, Lobna Osman, and Ahmed Elashry. Hgrbol2: human gait recognition for biometric application using bayesian optimization and extreme learning machine. *Future Generation Computer Systems*, 143:337–348, 2023.

[10] Thong Duy Nguyen and Milan Kresovic. A survey of top-down approaches for human pose estimation. *ArXiv*, abs/2202.02656, 2022.

[11] Zhengxiong Luo, Zhicheng Wang, Yan Huang, Liang Wang, Tieniu Tan, and Erjin Zhou. Rethinking the heatmap regression for bottom-up human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13264–13273, 2021.

[12] Vinicius F Rodrigues, Rodolfo S Antunes, Lucas A Seewald, Rodrigo Bazo, Eduardo S dos Reis, Uelison JL dos Santos, Rodrigo da R Righi, Luiz G da S Junior, Cristiano A da Costa, Felipe L Bertollo, et al. A multisensor architecture combining human pose estimation and real-time location systems for workflow monitoring on hybrid operating suites. *Future Generation Computer Systems*, 135:283–298, 2022.

[13] Lei Shi, Yimin Zhou, Juan Wang, Zuli Wang, Ding Chen, Haifeng Zhao, Wankou Yang, and Edward Szczerbicki. Compact global association based adaptive routing framework for personnel behavior understanding. *Future Generation Computer Systems*, 141:514–525, 2023.

[14] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference*, pages 740–755. Springer, 2014.

[15] Lin Zhao, Jie Xu, Shanshan Zhang, Chen Gong, Jian Yang, and Xinbo Gao. Perceiving heavily occluded human poses by assigning unbiased score. *Information Sciences*, 537:284–301, 2020.

[16] Lucas Stoffl, Maxime Vidal, and Alexander Mathis. End-to-end trainable multi-instance pose estimation with transformers. *arXiv preprint arXiv:2103.12115*, 2021.

[17] Thomas Golda, Tobias Kalb, Arne Schumann, and Jürgen Beyerer. Human pose estimation for real-world crowded scenarios. In *2019 16th IEEE international conference on advanced video and signal based surveillance (AVSS)*, pages 1–8. IEEE, 2019.

[18] Romain Guesdon, Carlos F Crispim-Junior, and Laure Tougne. Multitask metamodel for keypoint visibility prediction in human pose estimation. In *International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP)*, 2022.

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

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

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

[22] 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*, pages 5693–5703, 2019.

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

[24] Romain Guesdon, Carlos Crispim-Junior, and Laure Tougne. Dripe: A dataset for human pose estimation in real-world driving settings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2865–2874, 2021.

[25] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

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