PoseTED: A Novel Regression-Based Technique for Recognizing Multiple Pose Instances

Afsana Ahsan Jeny, Masum Shah Junayed, and Md Baharul Islam(✉)

Bahcesehir University, Yıldız, Çırağan Cd., 34349 Beşiktaş/Istanbul, Turkey
mdbaharul.islam@eng.bau.edu.tr

Abstract. Pose estimation for multiple people can be viewed as a hierarchical set predicting challenge. Algorithms are needed to classify all persons according to their physical components appropriately. Pose estimation methods are divided into two categories: (1) heatmap-based, (2) regression-based. Heatmap-based techniques are susceptible to various heuristic designs and are not end-to-end trainable, while regression-based methods involve fewer intermediary non-differentiable stages. This paper presents a novel regression-based multi-instance human pose recognition network called *PoseTED*. It utilizes the well-known object detector YOLOv4 for person detection, and the spatial transformer network (STN) used as a cropping filter. After that, we used a CNN-based backbone that extracts deep features and positional encoding with an encoder-decoder transformer applied for keypoint detection, solving the heuristic design problem before regression-based techniques and increasing overall performance. A prediction-based feed-forward network (FFN) is used to predict several key locations' posture as a group and display the body components as an output. Two available public datasets are tested in this experiment. Experimental results are shown on the COCO and MPII datasets, with an average precision (AP) of 73.7% on the COCO val. dataset, 72.7% on the COCO test dev. dataset, and 89.7% on the MPII datasets, respectively. These results are comparable to the state-of-the-art methods.

Keywords: Keypoints estimation · Pose recognition · Person detection · Transformer encoder-decoder · STN · FFN

1 Introduction

Humans frequently use various forms of gestures framed with different organs to communicate their cognitive and emotional wellness. Correctly recognizing those poses will aid in observing the present mental and physical state of the individual. Thus, human pose recognition is receiving research attention in recent years. The pose evaluation was necessary owing to the difficulty in identifying human joints in photos or films (often known as the essential points - knees, arms, elbows, etc.). In addition to being a challenging area of research, human pose measurement is critical in many practical uses,

This work is supported by the Scientific and Technological Research Council of Turkey (TUBITAK) under 2232 Outstanding Researchers program, Project No. 118C301.

G. Bebis et al. (Eds.): ISVC 2021, LNCS 13017, pp. 573–585, 2021.
https://doi.org/10.1007/978-3-030-90439-5_45

including action recognition, pedestrian detection, medical imaging, human-computer interaction, animation, health, and sports, and others. However, pose detection is a complicated subject that has remained unsolved. Many factors contribute to the difficulties, including prominent position and size diversity, cross-body connection, considerable visual variance, and background complexity. Several methods for estimating occluded joints involve using statistical and topological approaches to address the difficulties associated with occlusion [19].

To solve this problem, researchers have found solutions to this challenge by breaking it down into smaller tasks, including assessing a single-person stance, assessing multiple-person poses, and determining the pose of a human in a busy environment. To complete these sub-tasks, few methodologies have been tested in prior studies. First, top-down techniques utilize a bounding box as an object detector to identify and focus on a single individual to maximize efficiency [23]. Following that, calculating the posture of a single person, whereas with a bottom-up approach, examine various independent semantics elements and put them together to form a single-person pose [7]. Several heat map-based methods [23,30] are used to implement complex key points identification, which is then followed by subsequent processes for clustering and grouping. Only a few research studies [24,28,32] have focused on regression-based techniques, primarily because regression-based methods often perform worse than heat map-based methods in terms of accuracy and precision. However, heatmaps establish a gap between the overall estimate of the key points cannot make an end-to-end learning framework. On the other hand, regression-based methods are still inefficient because of a significant number of heuristic designs of the architecture, computation efficiency, and it is challenging to detect keypoint in occluded scenes [18,28].

Toshev et al. [26] introduced one of the earliest regression methods named DeepPose using AlexNet as a backbone, then extended it to learn essential key points from pictures. Since DeepPose has shown such remarkable results, the human pose estimation (HPE) research community has moved from traditional techniques to convolutional neural networks (CNNs) because of the weak localization. In 2017, a regression-based method, namely LCRNet, was proposed to identify human pose in both 2D and 3D. One disadvantage of this approach is that the restricted number of anchors places a constraint on the network's ability to estimate new positions. Luvizon et al. [15] combine soft-argmax function and soft-plus (which produces vectors with equivalent joint coordinates) to provide an utterly differentiable approach for HPE. However, the convolutional features are misaligned. It also faced typical localization and quantization problems. Then, in [16], a single pose regression-based method based on a transformer and CNN was presented. The model assumes a series of keypoint locations utilizing CNN and transformer simultaneously. This method is inefficient since it causes design issues and is computationally costly. However, the design of all recently advanced model architectures is more complicated since it has many levels of complexity.

Contributions. Motivated by the above observations, a regression-based 2D multi-instance human pose identification technique is proposed to utilize the modified YOLO-v4 for person detection. The keypoints detection transformer is used for pose estimation. For the feature extractor of the image of the bounding boxes, we utilize the CNN-based three networks such as DarkNet-53 [21], ResNet-101 [10], and VGG-19 [22] by eliminating the fully connected layer. As a result, the total number of parameters is minimized from 44.6 million, 143.6 million, and 40.5 million to 8.2 million, 25.8 million, and 5.6 million, respectively. Furthermore, to mitigate the parameters, the computing cost is significantly decreased as well. We compare the proposed architecture to the available regression-based methods to ensure the proposed method is more robust. It shows the competitive performance in posture recognition while eliminating the need for heuristic designs and allowing for quicker and more accurate pose recognition. The following contributions are significant in this paper.

- A novel regression-based multi-instance human pose recognition technique is proposed based on a general-purpose object detector, which is end-to-end trainable with two networks and can solve several limitations using transformers with YOLOv4.
- For faster and more accurate person detection, the modified YOLO-v4 [2] object detection model is used, followed by a cropping filter and passing images through a spatial transformer network (STN) to crop in the original images. The cropped images are fed into a CNN-based backbone that uses to extract covariance features. This feature passes into the transformer encoder-decoder using positional encoding, which involves defining the location of an object relative to its bounding box.
- DarkNet-53 [21], ResNet-101 [10], and VGG-19 [22] are three well-known CNN-based modified backbones are utilized for deep covariance feature extraction and decreased parameters. A prediction-based feed-forward network (FFN) was also employed to forecast the pose of multiple key points as a group and depict the body components as output.
- The efficacy of the PoseTED is experimentally shown on two complex benchmark datasets: the COCO [14] datasets (COCO val set and COCO test dev set), and the MPII [1] Human Pose dataset. It significantly improves the state-of-the-art performance on recent advanced regression-based techniques and is comparable to the heatmap-based methods.

2 Proposed PoseTED

Figure 1 depicts the PoseTED model's architecture. It has four parts: a modified YOLO-v4 well-known object detector is utilized for human detection with the bounding box; STN is used for cropping filter, then a CNN-based backbone network is utilized for feature extraction of the image; then it is connected to the transformer encoder-decoder combinations, and finally, feed-forward networks are used to detect and locate long-range spatial interactions of the linear combination of classes, person keypoint coordinates, and ultimately visible of identifying key points.

Person Detector (YOLO-v4). Our method is used for a regression approach to solve the multi-person pose recognition issue, and we used YOLO-v4 [2], a well-known object detection architecture, as the detection method. It is chosen for this study because of its excellent accuracy and reliability, the speed with which it can be assembled, simplicity of implementation, stability, and promise of acceptable results even in minute details. We studied the impact of various network resolutions, detection accuracy, and transfer learning parameters on detection outcomes while improving the YOLOv4 model. The updated YOLOv4 person detector is shown in Fig. 2.

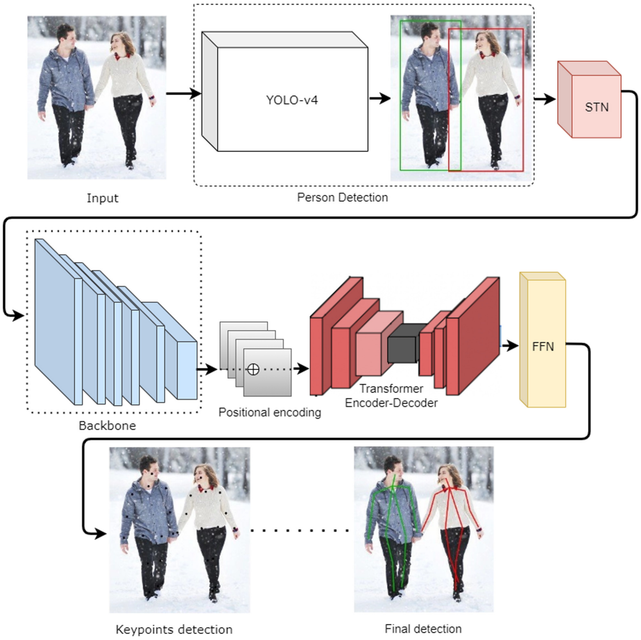

Fig. 1. The PoseTED architecture has been proposed. Here, YOLO-v4 is used for human detection with a bounding box. After detecting the person, the STN was utilized to crop the original image with the predicted boxes. Then, the backbone is used to generated in-depth features. The Transformer encoder-decoder with positional encoding works as keypoint-detection relative to corresponding bounding boxes. Finally, prediction-based FFN has predicted key points and displays vectors of body parts.

There are three main parts of this model: the spine, the neck, and the head. For reducing unwanted highlights, the spine is directly responsible. This CSPDarknet53 connection has been shown out to be an excellent choice. In [2], the yield is split into two portions: one is in the base layer and the other is in the reserved layer. One person heads to the Dense Block, while the other takes the following steps along the path as an exhibit presentation. Thick squares are composed of layers stacked on top of each other, with each subsequent layer beginning with Batch Normalization and ReLU, followed by a convolutional layer. A Dense Block layer is created by using all of the component guides from previous levels. That increases the area of the spinal column that may be accessed and aids in recognition of complicated image features. The concept of spatial pyramid pooling (SPP) [21] is used in neckbands to increase the acquiring field and provide interfaces that permit connections between different levels of the spine. A final portion has two parts; a classification head is placed on the object and assigned to the person or background detection, and a 4-channel regression head is applied for the predict and computed bounding boxes.

STN. After obtaining the detected pictures with bounding boxes, the STN [8] identifies the object as a person and predicts the bounding boxes to crop out the necessary portions

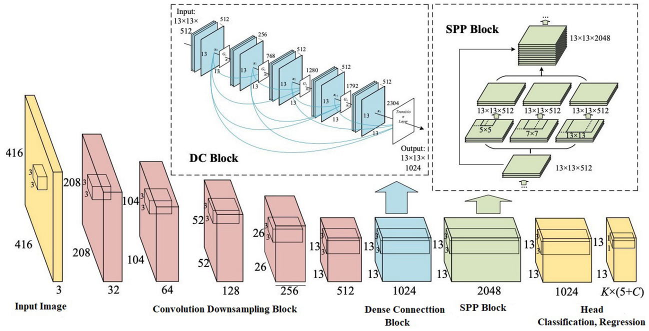

Fig. 2. The architecture of the YOLO-v4 person detector.

of the detected human with the bounding box. In the first step, STN selects the bounding box of the detected feature, rotates the area to normalize the posture, and then scales the cropped region before passing it to the CNN backbone. Feature maps and bounding box coordinates can be used to differentiate this cropping process. The grid $(w \times h)$ of the bounding box $(b = (x_l, x_r, y_t, y_d))$ is generated by,

$$x_i = \frac{w-i}{w} x_l + \frac{i}{w} x_r \tag{1}$$

$$y_j = \frac{h-j}{h} y_t + \frac{j}{h} y_d \tag{2}$$

where b, w and h are the original detected image's bounding box, height and weight of feature map respectively. And, (x_l, x_r, y_t) and y_d are denoted by coordinates of the bounding box in detected image respectively.

Backbone. The CNN architecture is often considered to be a backbone framework. The feature extractor in PoseTED's backbone extracts essential low-level characteristics such as body key points of humans. In order to arrive at more relevant comparisons, we have chosen three representative backbone designs, including ResNet, VGG, and DarkNet. Simplicity is essential when working with complex data sets such as images, and we thus maintain the first, pre-trained portion of the ImageNet CNN model as the initial layers to extract low-level characteristics from the input picture. After tuning these feature extractors, ResNet-101 [10], VGG-19 [22], and DarkNet-53 [21] are given to us as ResNet-P, VGG-P, and DarkNet-P. Compare to the three competitor backbone's parameters count are only 8.2 million, 25.8 million, and 5.6 million, respectively, representing 18.3%, 17.8%, and 13.8% of the original ResNet-101 (44.6 million), VGG-19 (143.6 million), and DarkNet-53 parameters, respectively (40.5 million).

Transformer Encoder Decoder. After getting the low-level feature maps of the CNN backbones, the detection of pose key points is used to the conventional architecture as the basis for the transformer encoder-decoder model with positional encoding [4]. There are six layers and eight attention heads in the encoder and decoder [27]. Reducing the size of the input channels by a 1×1 convolution is done in the encoder. The multi-head method requires sequential input; thus, the tensor is first downsampled and then compressed on the spatial dimension to one dimension. Because the transformer design is permutation-invariant, we encoded the picture as a fixed positional encoding to get the results we want. Instead, the positional encodings used by the decoder are known as object queries, and they correspond to the learned embeddings that we refer to as objects. Without permutation invariance, these queries will have to be changed when using the decoder because of the permutation. The decoder input is formed by adding them to the encoder output. The decoder first converts the queries to input embeddings before converting them to output embeddings. The poses are generated separately, and then the class labels and poses are decoded together to provide the final pose keypoints predictions and class labels. The resulting capability enables a query to find one instance, identify it, and forecast its posture keypoints and location.

Motivated by [5,12,34], we instead demand all decoder layers to anticipate the keypoint coordinates. The thing that we specifically allow is for the first decoder layer to guess the destination coordinates directly. Refining the predictions of each subsequent decoder layer helps improve the predictions of each preceding decoder layer. The key points then build on each other by becoming increasingly more nuanced.

FFN. After getting nuanced features from the transformer encoder-decoder, a 2-channel regression map predicts the coordinates of each key point. After that, the 3-layer perceptron with ReLU activation and a linear prediction layer estimates the final posture. The linear layer produces the class label using a softmax function, while the perceptron layer provides the center coordinate of key points. It produces an output consisting of the body component's respective coordinate and displacement values and the scores for every body part. The resulting vector is sent through a softmax classifier, which yields the class label. As a result, we made images normalize around their center and their offset from their origin.

3 Experiments

3.1 Dataset

MPII Dataset. There are around 25,000 photos of annotated body joints in the MPII [1] collection, covering more than 40,000 participants. The images were culled from YouTube videos depicting 410 different types of ordinary human activity. There are frames with 2D and 3D joint labels, head, torso alignments, body component occlusions in the collection, and frames with no labels. Another peculiarity of the MPII dataset is that it comprises frames from the prior and subsequent frames. However, they are not labeled.

COCO Dataset. It is the most broadly utilized enormous scope dataset. In our work, we used the COCO keypoints 2017 dataset. More than 200k pictures of the MS COCO keypoint dataset [14] in public, and more than 250k person instances are tagged with 17 human joints in this dataset. We train our networks using the COCO 2017 training dataset, which contains 57K pictures and 150K tagged human instances without using any additional data. We assess our networks' performance with the val2017 (5k pictures) and the test-dev2017 (20k images) datasets.

3.2 Experimental Setup and Training Details

To estimate human posture, we used a regression-based model. After detecting humans with the bounding box, cropped pictures of a single individual are provided in the training samples. All pictures are resized to a resolution of 384 × 288 and 256 × 192 pixels. As with [23], we used the same training methods and person's key detection outcomes. In the [31], they used to coordinate decoding method to minimize quantization error, which decodes detecting features from downsampled data, which is also used in our implementation. ReLU activates the function after training the feed-forward layers with 0.1 dropouts. Due to the ability to handle three backbones and the usage of the parameters, we configured the Transformer encoder hyperparameters in a way to avoid creating an excessive model capacity. To this end, thee stage networks are used throughout all of our studies. When throwing the forward pass, the time to execute is about 0.8–1.0 milliseconds on the Geforce RTX 2080 2080 GPU. When we compared the multi-view regression model's calculation time to the usage of the 2D detector, we can conclude that the computation time of our multi-view regression model is practically significant. For the COCO dataset, the Adam optimizer is used to train this model. To minimize error, first the learning rate at 0.001 is used, then exponentially decreased this rate to 0.0001. To conduct the validation of the MPII dataset, we randomly picked up 350 pictures from the training set to be used for this purpose. For the 300th epoch, all the single-person samples and the remaining training pictures are utilized. For extreme situations when centroids are overlapped, an offset is applied to significantly disrupt the centroids.

3.3 Evaluation Matrics

We utilized the well-accepted approach described in [23] and employed Object Keypoint Similarity (OKS) for the COCO dataset to assess the overall performance. For the MPII dataset, Percentage of Correct Keypoints (PCK) is used to determine the precision with which various keypoints may be located within a certain threshold. For each test image with the head segment length, the threshold is set to 50%, and it is marked as PCKh@0.5. The larger the PCK number, the greater the model's performance can be considered.

4 Results and Discussions

4.1 Performance on COCO Dataset

Table 1 shows the outcomes of several 2D pose detection techniques on the COCO dataset's validation dataset, as well as the description of the test parameters (approaches, backbones in models, input picture size, the size of parameters (PM), the GFLOPs network, and AP values). It is clear from Table 1 that our proposed approach PoseTED achieved more AP score than the popular TFPose (72.3 vs. 71) with the same input size 256 × 192. When the input size is increased to 384 × 288, our network outperformed others regression-based methods in terms of AP score (73.7%), particularly from DeepPose [26] (58.3%), PointSetNet [28] (65.7%), PointSetNet [28] (69.8%), and even from

Table 1. Comparing performance with other state-of-the-art studies using the COCO val. dataset. We compared the evaluation matrices to the backbones, input size, parameter size, and evaluation matrices.

Approaches	Backbone	Input size	PM	GFLOPs	AP	AP@50	AP@75
	Heatmap based approaches						
CMU-Pose [3]	VGG-19	-	-	-	64.2	86.2	70.1
Hourglass [17]	HourGlass-8	256 × 192	25.1M	14.3	66.9	-	-
CPN [6]	ResNfet-50	256 × 192	27M	6.20	68.6	-	-
SimpleBaseline [30]	ResNet-101	384 × 288	53M	26.7	73.6	89.6	80.3
HRNet [23]	HRNet-W32	384 × 288	28.5M	16.0	75.8	90.6	82.7
	Regression based approaches						
DeepPose [26]	ResNet-152	256 × 192	-	7.69	58.3	-	-
PointSetNet [28]	ResNeXt-101		-		65.7	85.4	71.8
PointSetNet [28]	HourGlass-W48	-	-	-	69.8	88.8	76.3
TFPose [16]	ResNet-50	256 × 192	-	9.2	71	-	-
TFPose [16]	ResNet-50	384 × 288	-	20.4	72.4	-	-
PoseTED	VGG-19	256 × 192	33.6M	13.8	68.2	88.7	78.6
PoseTED	ResNet-101	256 × 192	52.1M	19.0	70.5	89.1	78.9
PoseTED	DarkNet-53	256 × 192	52.8M	17.5	72.3	89.6	78.4
PoseTED	VGG-19	384 × 288	33.6M	17.0	71.4	90.1	80.3
PoseTED	ResNet-101	384 × 288	52.1M	23.2	72.1	90.3	80.5
PoseTED	DarkNet-53	384 × 288	52.8M	21.0	**73.7**	**90.5**	**80.9**

TFPose [16] (72.4%) as well as their backbone networks (ResNet-152, ResNeXt-101, and HourGlass-W48) are significantly larger than ours (DarkNet-53) except TFPose (ResNet-50). Furthermore, the PoseTED obtained higher AP scores than the famous heat map-based SimpleBaseline network (73.7 vs. 73.6) when the input size is the same as the SimpleBaseline network [30]. Moreover, in terms of 50 and 75% AP IoU thresholds, the performance of PoseTED is comparable with others.

Figure 3 depicts various qualitative findings from the COCO datasets, such as in the cases of walking, roller skating, basketball, exercising, inline skating, crossing, sitting, racing cars, and so on. It is seen from these instances that, although twisted postures have been observed among the public pictures, the estimates are still accurate enough. Identifying joints in limbs that are not adequately separated from one another and dislocate one another are examples of challenging circumstances.

Table 2 demonstrates the evaluation of the COCO test-dev set's performance with several regression-based methods such as DeepPose [26], CenterNet [33], Directpose [25], SPM [18], Integral [24], PointSetNet [28] and TFPose [16] are compared to our PoseTED as well as heat map-based approaches. As can be shown in Table 2, our PoseTED outperformed regression-based techniques such as DeepPose, Driectpose, and Integral while using the same backbone network and input size as the other two methods. Furthermore, it is also 3.3% and 0.4% higher than the heatmap-based methods such as G RMI [20] and Personlab [19], respectively, when the backbone network

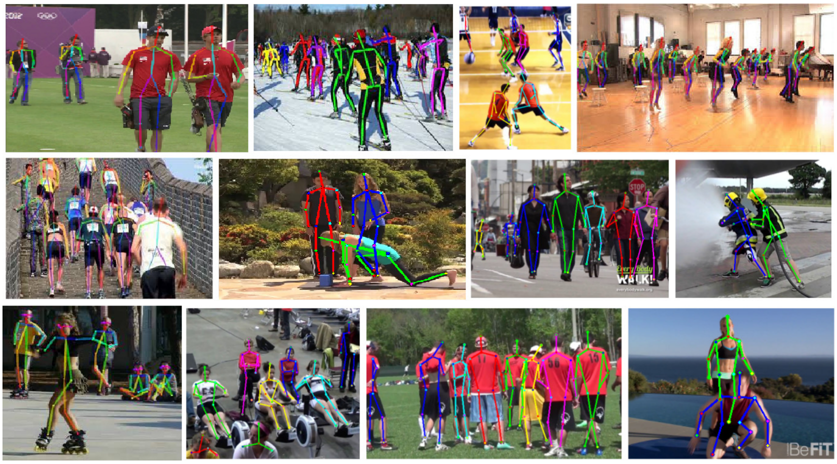

Fig. 3. Various qualitative results from the COCO datasets e.g. in the case of walking, roller skating, basketball, exercise, inline skatin, crossing, sitting, racing vehicles and so on.

Table 2. Comparing performance with other state-of-the-art studies using the COCO test dev. dataset. We compared the evaluation matrices to the backbones, input size, parameter size, and evaluation matrices.

Approaches	Backbone	Input size	PM	GFLOPs	AP	AP@50	AP@75
	Heatmap based approaches						
Mask RCNN [9]	ResNet-50	-	-	-	63.1	87.3	68.7
G RMI [20]	ResNet-101	353 × 256	42.6M	57.0	64.9	85.5	71.3
PifPaf [13]	Dilation ResNet-101	-	-	-	66.7	-	-
Personlab [19]	ResNet-101	-	-	-	67.8	88.6	74.4
Higher-HRNet [7]	HRNet-W48	-	-	-	70.5	89.3	77.2
DARK [31]	HRNet-W48	384 × 288	63.6M	32.9	76.2	92.5	83.6
	Regression based approaches						
DeepPose [26]	ResNet-101	256 × 192	-	7.69	57.4	86.5	64.2
CenterNet [33]	Hourglass-2	-	-	-	63.0	86.8	69.6
Directpose [25]	ResNet-101	-	-	-	63.3	86.7	69.4
SPM [18]	HourGlass	-	-	-	66.9	88.5	72.9
Integral [24]	ResNet-101	256 × 256	45.0M	11.0	67.8	88.2	74.8
PointSetNet [28]	HourGlass-W48	-	-	-	68.7	89.9	76.3
TFPose [16]	ResNet-50	256 × 192	-	9.2	70.9	90.5	79
TFPose [16]	ResNet-50	384 × 288	-	20.4	72.2	**90.9**	**80.1**
PoseTED	VGG-19	256 × 192	33.6M	13.8	67.3	88.8	74.7
PoseTED	ResNet-101	256 × 192	52.1M	19.0	68.2	89.4	75.8
PoseTED	DarkNet-53	256 × 192	52.8M	17.5	70.3	89.1	77.7
PoseTED	VGG-19	384 × 288	33.6M	17.0	69.1	90.3	76.5
PoseTED	ResNet-101	384 × 288	52.1M	23.2	69.8	90.6	77.5
PoseTED	DarkNet-53	384 × 288	52.8M	21.0	**72.7**	90.4	79.2

(ResNet-101) is the same. When the input size is extended to 384 × 288 pixels, the suggested PoseTED with the backbones (VGG-16 and ResNet-101) beat all regression-based techniques, except TFPose. However, when the DarkNet-53 is used as a backbone with the input size 384 × 288, the PoseTED obtained higher results even from TFPose (72.7 vs. 72.2). Furthermore, our network has greater GFLOPs than other networks, suggesting that it is also more efficient.

4.2 Performance on MPII Dataset

Our proposed model PoseTED has achieved the highest performance among the regression-based techniques when tested against the MPII dataset in Table 3. When VGG-19, ResNet-101, and DarkNet-53 are used as the backbone in our network, 89.7, 88.2, and 89.5 PCKh@0.5 scores have been obtained through our proposed model PoseTED which is higher than regression-based methods such as Integral [24] and Carreira et al. [5]. The PoseTED with the backbone network VGG-19 is also 2% higher than the heatmap-based method CPM [29] and 7.3% higher than the method presented by Hu et al. [11] (89.7% vs. 87.7% and 89.7% vs. 82.4%). PoseTED is generally comparable to heatmap-based techniques in terms of performance.

Table 3. Comparing performance with other state-of-the-art studies using the MPII dataset. We compared the evaluation matrices to backbones, and other evaluation matrices.

Approaches	Backbone	Head	Sho	Elb	Wri	Hip	Knee	Ank	Mean
	Heatmap based approaches								
Hu et al. [11]	VGG-16	95.0	91.6	83.0	76.6	81.9	74.5	69.5	82.4
CPM [29]	CPM	96.2	95.0	87.5	82.2	87.6	82.7	78.4	87.7
SBL [30]	ResNet-152	97.0	95.9	90.3	85.0	89.2	85.3	81.3	89.6
HRNet [23]	HRNet-W32	97.1	95.9	90.3	86.4	89.1	87.1	83.3	90.3
	Regression based approaches								
Carreira et al. [5]	-	95.7	91.7	81.7	72.4	82.8	73.2	66.4	81.3
Integral [24]	ResNet-101	-	-	-	-	-	-	-	87.3
PoseTED	VGG-19	96.4	94.9	88.4	82.6	**90.2**	84.1	78.4	**89.7**
PoseTED	ResNet-101	**97.9**	95.6	89.8	82.6	88.6	**85.4**	78.4	88.2
PoseTED	DarkNet-53	97.8	**96.0**	**90.0**	**84.3**	89.8	**85.2**	**79.7**	89.5

5 Conclusion

In this article, we introduced a new architecture for regression-based multi-instance pose recognition called PoseTED. It eliminates the need for complicated pre-processing and post-processing techniques and employs less heuristic approaches than the prior methods. The PoseTED employs YOLO-v4 well-known object detector for person detection with a bounding box and STN for cropping the original detected picture. Following cropping, three CNN-based backbones (DarkNet-53, VGG-19, and ResNet-101) are used to extract deep covariance low-level features. Then, a transformer encoder-decoder with positional encoding is utilized to match queries of human keypoints included in the loss calculation. The prediction-based FFN is then used to identify pose keypoints and visualize them as a vector between human body joints, resulting in enhanced performance. The experimental results on MS-COCO and MPII datasets are tested and compared to the recent advanced approaches. The PoseTED outperforms all contemporary techniques that are state of the arts. However, when people are highly obscured in situations, it is difficult to estimate these scenarios using our method in certain instances. Therefore, to be implemented in the future, we would like to work on the previously mentioned limitations and try to make more powerful backbone networks to experiment with regression-based person identification and posture recognition to enhance flexibility.

References

1. Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: 2D human pose estimation: new benchmark and state of the art analysis. In: Proceedings of the IEEE Conference on computer Vision and Pattern Recognition, pp. 3686–3693 (2014)
2. Bochkovskiy, A., Wang, C.Y., Liao, H.Y.M.: Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934 (2020)
3. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2D pose estimation using part affinity fields. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7291–7299 (2017)
4. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-End object detection with transformers. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12346, pp. 213–229. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58452-8_13
5. Carreira, J., Agrawal, P., Fragkiadaki, K., Malik, J.: Human pose estimation with iterative error feedback. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4733–4742 (2016)
6. Chen, Y., Wang, Z., Peng, Y., Zhang, Z., Yu, G., Sun, J.: Cascaded pyramid network for multi-person pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7103–7112 (2018)
7. Cheng, B., Xiao, B., Wang, J., Shi, H., Huang, T.S., Zhang, L.: Higherhrnet: scale-aware representation learning for bottom-up human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5386–5395 (2020)
8. Fang, Y., Zhan, B., Cai, W., Gao, S., Hu, B.: Locality-constrained spatial transformer network for video crowd counting. In: 2019 IEEE International Conference on Multimedia and Expo (ICME), pp. 814–819. IEEE (2019)
9. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2961–2969 (2017)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
11. Hu, P., Ramanan, D.: Bottom-up and top-down reasoning with hierarchical rectified gaussians. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5600–5609 (2016)

12. Hu, T., Qi, H., Xu, J., Huang, Q.: Facial landmarks detection by self-iterative regression based landmarks-attention network. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
13. Kreiss, S., Bertoni, L., Alahi, A.: Pifpaf: Composite fields for human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11977–11986 (2019)
14. Lin, T.Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
15. Luvizon, D.C., Tabia, H., Picard, D.: Human pose regression by combining indirect part detection and contextual information. Comput. Graph. **85**, 15–22 (2019)
16. Mao, W., Ge, Y., Shen, C., Tian, Z., Wang, X., Wang, Z.: Tfpose: Direct human pose estimation with transformers. arXiv preprint arXiv:2103.15320 (2021)
17. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9912, pp. 483–499. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46484-8_29
18. Nie, X., Feng, J., Zhang, J., Yan, S.: Single-stage multi-person pose machines. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6951–6960 (2019)
19. Papandreou, G., Zhu, T., Chen, L.C., Gidaris, S., Tompson, J., Murphy, K.: Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 269–286 (2018)
20. Papandreou, G., et al.: Towards accurate multi-person pose estimation in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4903–4911 (2017)
21. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767 (2018)
22. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
23. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5693–5703 (2019)
24. Sun, X., Xiao, B., Wei, F., Liang, S., Wei, Y.: Integral human pose regression. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 529–545 (2018)
25. Tian, Z., Chen, H., Shen, C.: Directpose: Direct end-to-end multi-person pose estimation. arXiv preprint arXiv:1911.07451 (2019)
26. Toshev, A., Szegedy, C.: Deeppose: Human pose estimation via deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1653–1660 (2014)
27. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 5998–6008 (2017)
28. Wei, F., Sun, X., Li, H., Wang, J., Lin, S.: Point-set anchors for object detection, instance segmentation and pose estimation. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12355, pp. 527–544. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58607-2_31
29. Wei, S.E., Ramakrishna, V., Kanade, T., Sheikh, Y.: Convolutional pose machines. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4724–4732 (2016)
30. Xiao, B., Wu, H., Wei, Y.: Simple baselines for human pose estimation and tracking. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 466–481 (2018)

31. Zhang, F., Zhu, X., Dai, H., Ye, M., Zhu, C.: Distribution-aware coordinate representation for human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7093–7102 (2020)
32. Zhou, X., Koltun, V., Krähenbühl, P.: Tracking objects as points. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12349, pp. 474–490. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58548-8_28
33. Zhou, X., Wang, D., Krähenbühl, P.: Objects as points. arXiv preprint arXiv:1904.07850 (2019)
34. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159 (2020)