LIGHTPOSE: A LIGHTWEIGHT AND EFFICIENT MODEL WITH TRANSFORMER FOR HUMAN POSE ESTIMATION

Xiyang Liu, Peng Li, Ding Ni, Yan Wang, Hui Xue*

Alibaba Group

ABSTRACT

The prediction of keypoints by generating high-resolution heatmaps has become a popular solution in human pose estimation. While this kind of method requires up-sampling or deconvolution operations, which would bring a great challenge to the acceleration of model inference. If performing keypoint prediction on low-resolution heatmaps, the performance is unsatisfied due to serious quantization errors. To solve this contradiction, we propose to perform joint training of the heatmap and center offset on low-resolution heatmaps to reduce quantization errors, which could achieve the comparable performance to the high-resolution heatmap and reduce the computational complexity. In addition, we utilize transformer to enhance the representation ability of low-resolution features, instead of increasing the network layers or the convolution kernel size. The transformer could bring the significant improvement of the performance with little computation cost. Combining the above two modules, we design a new lightweight pose estimation model, named LightPose. Experimental results have shown that, compared with HRNet, our method could achieve the state-of-the-art performance on COCO and MPII datasets with a massive reduction of the parameters by 86% and GFLOPs by 67%.

***Index Terms*—** Human Pose Estimation, Lightweight Model, Visual Transformer

1. INTRODUCTION

Due to the occlusion and complex changes of human pose, accuracy keypoint estimation is still a hard task in computer vision. How to represent the keypoint positions of human body and provide strong supervision for training is an open problem that has been explored in this field. On the one hand, DeepPose [1] and others [2, 3] directly use the keypoint positions as training targets. This kind of supervising method outputs the coordinate information of keypoints directly. While the spatial location information is lost, and the generalization ability is not good enough to achieve satisfactory performance. On the other hand, many works [4, 5, 6, 7, 8] adopt heatmap as the training target in keypoint estimation. The principle of heatmap is to generate 2D Gaussian responses in the center of keypoints, so the spatial information is retained in the training phase. Unfortunately, the heatmap has a problem of quantization error, and the lower resolution of it is, the larger error it has. Therefore, SimpleBase [9], HRNet [10] and HigherHRNet [11] predict the heatmaps with higher resolution to keep more details and alleviate quantization errors. But the drawback of high-resolution prediction is also obvious, the calculation cost is greatly increased.

In order to achieve a good balance between the accuracy and the calculation cost, we propose a low-resolution supervising label based on the heatmap and the center offset (local region deviations of each ground-truth keypoint). Different from previous works [12, 13], we generate Gaussian heatmaps for rough positioning and assign different weights to the center and its surrounding points to learn accurate quantization errors of low-resolution heatmaps for fine positioning. As shown in Figure 1, the performance of training with the heatmap would drop sharply with the reduction of output resolution. While our method with joint training of heatmap and center offset is insensitive to the feature resolution and performs well on low-resolution. The low-resolution supervising method could avoid the operations of up-sampling or deconvolution, which greatly reduces the computational complexity of the model.

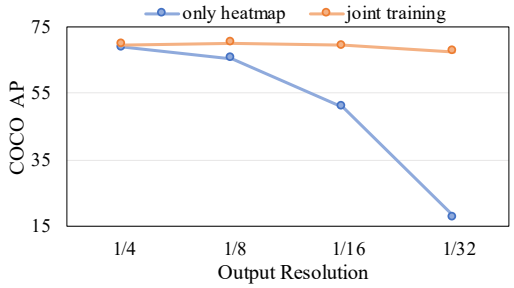

Fig. 1. The effect of different output resolutions on two kinds of supervising methods with the same backbones, including training only with the heatmap and joint training of the heatmap and the center offset.

Furthermore, in order to enhance the feature representation ability of the model at low-resolution, we utilize the transformer [14] to obtain global context in human pose estimation. The transformer can capture long distance interactions through the self-attention between any two pixels, and it is more effective than Convolutional Neural Network (CNN) with limited receptive field. The transformer has shown great potential on the main tasks of computer vision, such as object classification [15, 16], object detection [17, 18], semantic segmentation[19] and so on [20, 21]. Specifically, we apply multiple transformer encoders on high-level features with low-resolution, so the length of transformer sequences is short and the computation cost is little. Integrating the transformer into our proposed network by this way makes the model more efficient while keeping it lightweight. The similar network design method is also applicable to other single person pose estimation networks.

In summary, our contributions are as follows:

(1) We propose to perform joint training of Gaussian heatmap and center offset on low-resolution features, which could obtain comparable performance to the high-resolution heatmap and greatly reduce the computational complexity.

(2) We introduce the transformer into keypoint estimation network to enrich the global information and we provide a new lightweight model design paradigm for the community.

*Contacted e-mail: xiyang.lxy@alibaba-inc.com

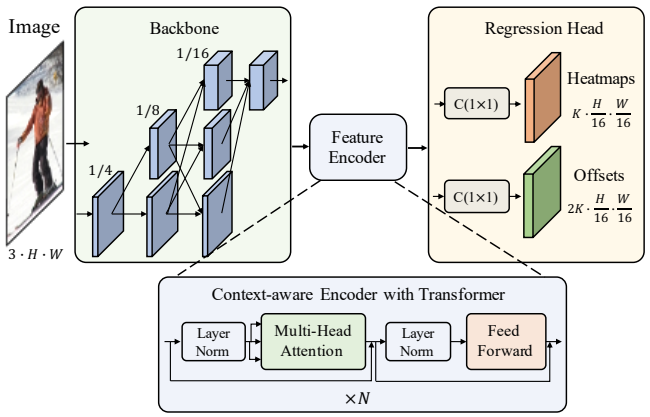

Fig. 2. The overview of our approach mainly includes three modules: 1) Backbone: extracting a high-level low-resolution feature map based on adjusted HRNet or ResNet. 2) Feature Encoder: flatting the feature map into a sequence and capturing the context information by multiple transformer encoders. 3) Regression Head: It consists of two 1x1 convolution layers, and learns the Gaussian response of each position to locate a keypoint and the center point offset of each one to obtain a refined result, respectively.

(3) Experiments demonstrate that the proposed method is superior to other complex state-of-the-art methods in terms of model parameters, calculation cost and prediction accuracy on the typical datasets.

2. PROPOSED METHOD

In this section, we introduce the low-resolution supervising label and context-aware encoder with transformer in details. Based on above modules, we design a new lightweight model named LightPose for human pose estimation. The overview of LightPose is depicted in Figure 2 and its two detailed network structures including LightPose-R50 and W48 are shown in Table 1.

2.1. Low-Resolution Supervising Label

Heatmap Learning. For a keypoint with the ground truth location p, its corresponding low-resolution location is $\widetilde{p} = \lfloor \frac{p}{R} \rfloor$, where R is the down-sample factor. Given an $W \times H \times 3$ image containing K keypoints, we generally splat all keypoints onto K maps. Each keypoint map is smoothed by a Gaussian kernel $G_\sigma = \exp(-\frac{(x-\widetilde{p}_x)^2+(y-\widetilde{p}_y)^2}{2\sigma})$ at the low-coordinate position $\widetilde{p}$. Then the heatmap $H_k \in [0,1]^{\frac{W}{R} \times \frac{H}{R}}$ of the k-th keypoint is represented as $H_k = \delta(x - \widetilde{p}_x, y - \widetilde{p}_y) \cdot G_\sigma$.

Generally, we optimize the MSE loss between the targeted heatmap H_k and the predicted heatmap $\hat{H}_k$ as follows:

$$L_{hm} = \frac{1}{K} \sum_{k=1}^{K} \left\| H_k - \hat{H}_k \right\|_2^2. \tag{1}$$

Actually, when increasing the down-sampling factor R, the quantization error of the heatmap would become serious. When R is set to a small value, it means that more operations are needed to restore the low-resolution high-level features to the high-resolution. While our experiments in Table 5 show that generating the high-resolution heatmap is not the only solution, and the joint learning of heatmap and center offset could achieve the same performance with little computation cost. Therefore, we introduce the center offset to reduce quantization errors in low-resolution heatmaps.

Table 1. The detailed network structures of LightPose-R50 (based on ResNet-50) and LightPose-W48 (based on HRNet-W48).

Architecture	LightPose-R50	LightPose-W48
Backbone	Conv-k7-s2-c64, BN, ReLU	Conv-k3-s2-c64, BN, ReLU
	Max Pooling-k3-s2	Conv-k3-s2-c64, BN, ReLU
	3×Bottleneck-c64	4×Bottleneck-c64
	4×Bottleneck-c128	Transition1-Stage2
	6×Bottleneck-c256	Transition2-Stage3
	Conv-k1-s1-c256	Conv-k1-s1-c256
Encoder	12×Transformer Encoder-d_m256-h8-d_o512	
Head	heatmap: Conv-k1-s1-c(K), offset: Conv-k1-s1-c(2K)	

Center Offset Learning. For the k th keypoint, we generate the center offset map $O_k(p) = |\frac{p}{R} - \widetilde{p}| \in R^{\frac{W}{R} \times \frac{H}{R} \times 2}$, in which each value represents the quantization distance from this point to the coordinates of the ground truth keypoint [13].

In fact, the location with the largest response value in the predicted heatmap may not be the real center point, so the points around the real center point also need to be paid more attention. Therefore, to further ensure the network capable of focusing on foreground pixels (the real center point) and hard background pixels (the points around the real center point), we define the center offset weight W_k to be:

$$W_k(p) = \begin{cases} H_k(p) & \text{if } H_k(p) > t \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

where the weight W_k has the same shape as the heatmap H_k, and threshold t is set to 0.1 in our experiments. The loss of the center offset is optimized with L1 distance:

$$L_{off} = \frac{1}{K} \sum_{k=1}^{K} W_k \cdot \left| O_k - \hat{O}_k \right|. \tag{3}$$

Finally, we formulate the overall loss function for low-resolution supervision during training as:

$$L_{all} = L_{hm} + \alpha L_{off}, \tag{4}$$

where α is the balancing weight between the two loss terms and set to 1.0 in our experiments.

2.2. Context-aware Encoder with Transformer

Due to the limited receptive field of CNN, the employed low-resolution feature which lacks global context information, would hinder further improvement of the performance. Therefore, we utilize the combination of multiple transformer encoders to replace the deeper blocks of CNN backbone (the c5 block in ResNet and the stage4 block in HRNet) as the context-aware encoder, which could greatly reduce the computational complexity compared with the pure CNN network.

The structures of context-aware encoder with transformer is shown in Figure 2, which mainly includes layer normalization, multi-head attention and feed-forward network. Layer normalization is always applied before multi-head attention and feed-forward network following [26]. Multi-head attention comprises multiple self-attention layers and it could encapsulate multiple complex relationships among different positions in a sequence. It could be calculated as follows:

$$Z = softmax(\frac{QK^T}{\sqrt{d}}) \cdot V, \tag{5}$$

where Q, K and V are three different transformation matrices of the input sequence X.

Table 2. Comparisons with typical methods including SimpleBase and HRNet on the COCO validation set.

Method	Backbone	Input size	Output ratio	#Params	FLOPs	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
SimpleBase [9]	ResNet-50	256 × 192	1/4	34.0 M	8.9 G	70.4	88.6	78.3	67.1	77.2	76.3
SimpleBase [9]	ResNet-152	256 × 192	1/4	68.6 M	15.7 G	72.0	89.3	79.8	68.7	78.9	77.8
LightPose-R50 (**ours**)	ResNet-50	256 × 192	1/16	**12.0 M**	**3.9 G**	**72.8**	**89.5**	**80.1**	**69.4**	**80.0**	**78.6**
HRNet-W32 [10]	HRNet-W32	256 × 192	1/4	28.5 M	7.1 G	74.4	90.5	81.9	70.8	81.0	79.8
HRNet-W48 [10]	HRNet-W48	256 × 192	1/4	63.6 M	14.6 G	**75.1**	**90.6**	**82.2**	**71.5**	**81.8**	**80.4**
LightPose-W48 (**ours**)	HRNet-W48	256 × 192	1/16	**8.8 M**	**4.9 G**	74.3	89.7	81.2	70.6	81.5	79.8

Table 3. Comparisons with state-of-the-art methods on the COCO test-dev set. #Params and FLOPs are calculated for the pose estimation network.

Method	Backbone	Input size	#Params	FLOPs	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
G-RMI [12]	ResNet-101	353 × 257	42.6 M	57.0 G	64.9	85.5	71.3	62.3	70.0	69.7
Integral Pose Regression [3]	ResNet-101	256 × 256	45.0 M	11.0 G	67.8	88.2	74.8	63.9	74.0	–
RMPE [22]	PyraNet [6]	320 × 256	28.1 M	26.7 G	72.3	89.2	79.1	68.0	78.6	–
CPN [23]	ResNet-Inception	384 × 288	–	–	72.1	91.4	80.0	68.7	77.2	78.5
CPN (ensemble) [23]	ResNet-Inception	384 × 288	–	–	73.0	91.7	80.9	69.5	78.1	79.0
SimpleBase [9]	ResNet-152	384 × 288	68.6 M	35.3 G	73.7	91.9	81.1	70.3	80.0	79.0
HRNet-W32 [10]	HRNet-W32	384 × 288	28.5 M	16.0 G	74.9	92.5	82.8	71.3	80.9	80.1
HRNet-W48 [10]	HRNet-W48	384 × 288	63.6 M	32.9 G	75.5	92.5	83.3	71.9	81.5	80.5
MSPN [24]	ResNet-50 ×4	384 × 288	120.3 M	53.6 G	76.1	**93.4**	**83.8**	72.3	81.5	**81.6**
DARK [25]	HRNet-W48	384 × 288	63.6 M	32.9 G	**76.2**	92.5	83.6	**72.5**	**82.4**	81.1
LightPose-R50 (**ours**)	ResNet-50	384 × 288	12.0 M	**8.7 G**	74.7	92.2	82.6	71.1	80.8	79.9
LightPose-W48 (**ours**)	HRNet-W48	384 × 288	**8.8 M**	11.0 G	75.1	92.2	82.6	71.5	81.2	80.2

A feed-forward network is applied after multi-head attention layers in each encoder with two linear layers and a ReLU activation function. It is denoted as the following function:

$$FFN(X) = W_2 \cdot ReLU(W_1 X), \tag{6}$$

where W_1 and W_2 are the two parameter matrices of the two linear layers.

3. EXPERIMENTS

3.1. Datasets

We evaluate our method on two publicly human pose estimation benchmark datasets: COCO2017 [27] and MPII [28]. The COCO2017 dataset contains over $250K$ person instances labeled with 17 keypoints, where $57K$ images are used for training, $5K$ images for validation and $20K$ images for testing. The MPII dataset contains of $25K$ images with $40K$ subjects, labeled with 16 keypoints for each.

3.2. Implementation Details

Evaluation Metric. In the COCO dataset, we report standard average precision and recall scores: AP, AP^{50}, AP^{75}, AP^{M} for medium objects, AP^{L} for large objects, and AR. The standard metric of the MPII dataset is the PCKh score, and PCKh@0.5 score is reported in our experiments.

Data Augmentation. For fair comparison, we adopt the same data augmentation strategies as HRNet [10], including random rotation ($[45°, 45°]$), random scale ($[0.65, 1.35]$), flipping and half body data augmentation.

Training Details. Following previous works, we extend the human detection box to a fixed aspect ratio: $height : width = 4 : 3$, and then crop the box from the original image, which is resized to a fixed size, 256×192 or 384×288. We use the Adam optimizer [29] and the training process is terminated within 220 epochs. The base learning rate is set as 1e-3, and is dropped to 1e-4 and 1e-5 at the 160th and 200th epochs, respectively.

3.3. Experimental Results

Results on the COCO validation set. We report the results of our method and other typical methods (SimpleBase [9] and HRNet [10]) in Table 2. As shown in the first group, our LightPose-R50 achieves the 72.8 AP score, outperforming SimpleBase with all backbones. Compared with Simple Baseline taking ResNet50 as the backbone, our LightPose-R50 improves 2.4 points in AP with 35.3% parameters and 43.8% GFLOPs. When it comes to a backbone of ResNet152, our LightPose-R50 could further improve 0.8 points in AP only with 17.5% parameters and 24.8% GFLOPs. As shown in the second group, compared with HRNet-W32, our LightPose-W48 could achieve an approximate performance with 30.9% parameters and 69.0% GFLOPs. Our Lightpose-W48 is only 0.8 points lower than HRNet-W48, while the parameters and GFLOPs are significantly reduced to 13.8% and 33.6%, respectively.

Results on the COCO test-dev set. Table 3 reports the pose estimation performances of our approach and the existing state-of-the-art approaches. Our approach has absolute advantages in terms of the model size and computational complexity. LightPose-R50 has the minimal computational cost only with 8.7 GFLOPs and obtains 74.7 AP. Moreover, LightPose-W48 has the minimal model parameters only with 8.8M parameters and could achieve 75.1 AP.

Results on the MPII validation set. Table 4 shows the PCKh@0.5 results, the model parameters and the GFLOPs of the top-performed methods. Our LightPose-R50 and LightPose-W48 achieve 89.6 and 90.6 PKCh@0.5 score, respectively, which outperform other typical methods such as Hourglass[4], SimpleBase [9] and HRNet [10]. In addition, our approach is also outstanding in term of the model size and the calculation cost. Compared with SimpleBase, LightPose-R50 only needs 17.5% of its parameters and 24.9% of its computation to achieve the same performance. Compared with HRNet-W32 and DARK, LightPose-W48 achieves better performance and reduces the model parameters and GFLOPs to 30.9% and 68.4%, respectively.

Fig. 3. Qualitative results of our method for some example images from the COCO (top) and MPII (bottom) datasets: containing viewpoint and appearance change, occlusion, multiple persons, and common imaging artifacts.

Table 4. Performance comparisons on the MPII validation set (PCKh@0.5)

Method	#Params	FLOPs	PCKh@0.5
Fast Pose [7]	**3.0 M**	9.0 G	89.0
Hourglass [4]	25.1 M	19.1 G	89.2
LFP [6]	28.1 M	21.3 G	89.6
SimpleBase [9]	68.6 M	20.9 G	89.6
DLCM [30]	15.5 M	15.6 G	89.8
HRNet-W32 [10]	28.5 M	9.5 G	90.3
DARK [25]	28.5 M	9.5 G	**90.6**
LightPose-R50 (**ours**)	12.0 M	**5.2 G**	89.6
LightPose-W48 (**ours**)	8.8 M	6.5 G	**90.6**

3.4. Ablation Studies

Resolution and Supervising Label. We explore the impact of output resolution on different training targets, as shown in Table 5. First of all, as shown in the first group of comparative experiments, when we only use the heatmap as the training label, the AP will sharply decrease from 69.0 to 18.0 with the decrease of the output feature map resolution from $1/4$ to $1/32$. It indicates that the prediction result of the heatmap is very sensitive to the resolution. Therefore, the previous works are more inclined to predict the heatmap with higher resolution. Secondly, as shown in the second group of comparative experiments, when we adopt both the heatmap and the offset as training labels, the prediction results of AP are very close, and the maximum error is only 0.8 between $1/4$, $1/8$ and $1/16$ resolutions. When the resolution is $1/32$, the performance begins to decrease obviously and the ΔAP is -2.1, due to the absolute size is only 6×8. The experiment indicates that high-resolution feature map is not a necessary condition for accurate keypoint estimation and joint training of heatmap and offset on low-resolution feature also works. For balance, we adopt the heatmap and the offset with $1/16$ resolution as the low-resolution supervising label.

Component Analysis. To better analyse the gain of the proposed components, we perform detailed ablation studies on each individual component are shown in Table 6. The combination of backbone and heatmap can only obtain a baseline of 51.0 AP, due to the quantization error of the low resolution heatmap. The backbone with the joint training of heatmap and center offset, could achieve a significant performance improvement to 69.3 AP. It indicates that the center offset makes up for the quantization error of low resolution heatmap. After that, the backbone followed by the transformer encoder could provide features that integrate context information for more accurate keypoint estimation, and the performance is further improved to 74.3 AP with only 3.2 M parameters and 0.6 GFLOPs calculation cost increasing.

Table 5. The influence of resolution on different supervising labels. ΔAP represents the AP error compared with the best result.

Output ratio	Supervising labels	AP	AR	Δ AP
1/4	heatmap	**69.0**	**72.3**	0.0
1/8	heatmap	65.5	69.1	-3.5
1/16	heatmap	51.0	56.8	-17.0
1/32	heatmap	18.0	26.4	-51.0
1/4	heatmap + offset	69.6	72.9	-0.5
1/8	heatmap + offset	**70.1**	**73.1**	0.0
1/16	heatmap + offset	69.3	72.4	-0.8
1/32	heatmap + offset	68.0	71.0	-2.1

Table 6. Ablation study on the components of LightPose. The resolutions of heatmap and center offset are 1/16 of the input image size.

Backbone	Encoder	Heatmap	Offset	AP	#Params	FLOPs
✓		✓		51.0	5.6 M	4.3 G
✓	✓	✓		54.8	8.8 M	4.9 G
✓		✓	✓	69.3	5.6 M	4.3 G
✓	✓	✓	✓	74.3	8.8 M	4.9 G

4. CONCLUSION

In this paper, we propose to perform joint training of the heatmap and the center offset on low-resolution features, which could obtain comparable performance to the high-resolution heatmap and greatly reduce the computational complexity of the model. Meanwhile, we introduce the transformer encoder instead of increasing the depth of CNN to enrich global context with little computation cost. Through above works, we provides a new lightweight model design paradigm for human pose estimation task. Experiments show that our proposed model LightPose has less computation and memory than other complex representative models, and obtains the state-of-the-art performance on the typical datasets.

5. REFERENCES

[1] Alexander Toshev and Christian Szegedy, "Deeppose: Human pose estimation via deep neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[2] Xiao Sun, Jiaxiang Shang, Shuang Liang, and Yichen Wei, "Compositional human pose regression," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[3] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei, "Integral human pose regression," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[4] Alejandro Newell, Kaiyu Yang, and Jia Deng, "Stacked hourglass networks for human pose estimation," in *European Conference on Computer Vision*, 2016.

[5] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh, "Convolutional pose machines," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[6] Wei Yang, Shuang Li, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang, "Learning feature pyramids for human pose estimation," in *IEEE International Conference on Computer Vision*, 2017.

[7] Feng Zhang, Xiatian Zhu, and Mao Ye, "Fast human pose estimation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[8] Yuanhao Cai, Zhicheng Wang, Zhengxiong Luo, Binyi Yin, Angang Du, Haoqian Wang, Xinyu Zhou, Erjin Zhou, Xiangyu Zhang, and Jian Sun, "Learning delicate local representations for multi-person pose estimation.," in *European Conference on Computer Vision*, 2020.

[9] Bin Xiao, Haiping Wu, and Yichen Wei, "Simple baselines for human pose estimation and tracking," in *European Conference on Computer Vision*, 2018.

[10] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang, "Deep high-resolution representation learning for human pose estimation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[11] Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S Huang, and Lei Zhang, "Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[12] George Papandreou, Tyler Zhu, Nori Kanazawa, Alexander Toshev, Jonathan Tompson, Chris Bregler, and Kevin Murphy, "Towards accurate multi-person pose estimation in the wild," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[13] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl, "Objects as points," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017.

[15] Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Masayoshi Tomizuka, Kurt Keutzer, and Peter Vajda, "Visual transformers: Token-based image representation and processing for computer vision," in *arXiv preprint arXiv:2006.03677*, 2020.

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, and Neil Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.

[17] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko, "End-to-end object detection with transformers," in *European Conference on Computer Vision*, 2020.

[18] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai, "Deformable detr: Deformable transformers for end-to-end object detection," in *International Conference on Learning Representations*, 2021.

[19] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip H. S. Torr, and Li Zhang, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," *arXiv preprint arXiv:2012.15840*, 2020.

[20] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo, "Learning texture transformer network for image super-resolution," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[21] Ruijin Liu, Zejian Yuan, Tie Liu, and Zhiliang Xiong, "End-to-end lane shape prediction with transformers," in *IEEE Winter Conference on Applications of Computer*, 2020.

[22] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu, "Rmpe: Regional multi-person pose estimation," in *IEEE International Conference on Computer Vision*, 2017.

[23] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun, "Cascaded pyramid network for multi-person pose estimation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[24] 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.

[25] Feng Zhang, Xiatian Zhu, Hanbin Dai, Mao Ye, and Ce Zhu, "Distribution-aware coordinate representation for human pose estimation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[26] Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tieyan Liu, "On layer normalization in the transformer architecture," in *International Conference on Machine Learning*, 2020.

[27] 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 *European Conference on Computer Vision*, 2014.

[28] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele, "2d human pose estimation: New benchmark and state of the art analysis," in *IEEE Conference on computer Vision and Pattern Recognition*, 2014.

[29] Diederik P. Kingma and Jimmy Lei Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations*, 2015.

[30] Wei Tang, Pei Yu, and Ying Wu, "Deeply learned compositional models for human pose estimation," in *European Conference on Computer Vision*, 2018.