

# Location-Free Human Pose Estimation

Xixia Xu<sup>1\*</sup>, Yingguo Gao<sup>2</sup>, Ke Yan<sup>2†</sup>, Xue Lin<sup>1</sup>, Qi Zou<sup>1†</sup>

<sup>1</sup>Beijing Key Laboratory of Traffic Data Analysis and Mining, Beijing Jiaotong University, China

<sup>2</sup>Tencent Youtu Lab, Shanghai, China

{19112036, 18112028, qzou}@bjtu.edu.cn; {yingguogao, kerwinyan}@tencent.com

## Abstract

Human pose estimation (HPE) usually requires large-scale training data to reach high performance. However, it is rather time-consuming to collect high-quality and fine-grained annotations for human body. To alleviate this issue, we revisit HPE and propose a location-free framework without supervision of keypoint locations. We reformulate the regression-based HPE from the perspective of classification. Inspired by the CAM-based weakly-supervised object localization, we observe that the coarse keypoint locations can be acquired through the part-aware CAMs but unsatisfactory due to the gap between the fine-grained HPE and the object-level localization. To this end, we propose a customized transformer framework to mine the fine-grained representation of human context, equipped with the structural relation to capture subtle differences among keypoints. Concretely, we design a Multi-scale Spatial-guided Context Encoder to fully capture the global human context while focusing on the part-aware regions and a Relation-encoded Pose Prototype Generation module to encode the structural relations. All these works together for strengthening the weak supervision from image-level category labels on locations. Our model achieves competitive performance on three datasets when only supervised at a category-level and importantly, it can achieve comparable results with fully-supervised methods with only 25% location labels on MS-COCO and MPII.

## 1. Introduction

Human pose estimation (a.k.a., keypoint localization) is a challenging yet fundamental computer vision task, which aims to detect the keypoint locations (e.g., eyes, ankles). In recent years, HPE has witnessed dramatic progress with the development of CNNs. An integral factor of the achievement is the availability of large-scale training data with

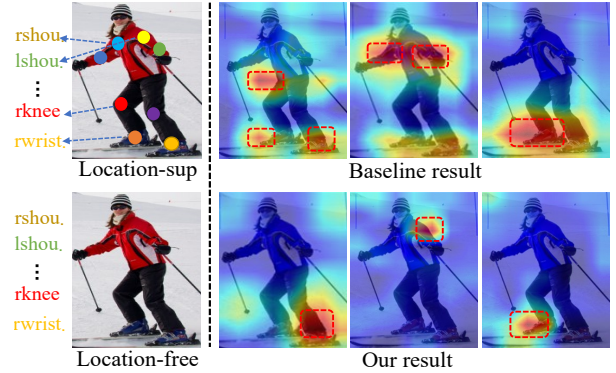


Figure 1. Column 1: **Location-supervised v.s. Location-free** HPE; Column 2-4: the result of location-free baseline v.s. the result of our method. Noted that the correct joints are lankle, lshou and rankle respectively.

precise location annotations. However, it's rather label-intensive and time-consuming to collect high-quality and fine-grained annotations. Thus, we study the keypoint localization when only the image-level category labels are given.

The Class Activation Mapping (CAM) [53] is a simple but effective method to discover object regions from intermediate classifier activation with only image-level labels, which has been the cornerstone of weakly-supervised object localization (WSOL) [53] and weakly-supervised semantic segmentation (WSSS) [47]. The CAM tends to focus on the most discriminative parts of object, and many approaches [20, 46, 48] are proposed to improve CAM to cover the full extend of an object. But these methods cannot localize the subtle joints due to the gap between the object-level localization and fine-grained keypoint localization.

Our method is also built upon CAM. There are two main obstacles for applying CAM in HPE. **i)** For tiny and local keypoints, it's hard for the model to capture the precise spatial features for accurate prediction without the explicit location labels. Furthermore, the local appearance features learned only relying on the image-level labels are not comprehensive enough for understanding the human body. Therefore, more contextual information and explicit spatial

\*Works done while interning at Tencent Youtu Lab.

†Corresponding author.

prior need to be exploited. **ii)** The inter-class differences among joints are subtle and the adjacent or symmetric joints possess similar semantic context or appearances. It tends to result in the location confusions and incorrect responses as in Fig. 1 (Baseline result in column 2-4). It’s hard for model to capture the fine-grained joint-specific features to eliminate the confusions especially without the explicit location supervision. The inherent structural relation between joints plays a critical role in helping distinguish or infer the uncertain locations. Thus, how to dig out the intrinsic structural relation prior for the model is vital.

Based on above discussions, we propose a novel customized transformer-based architecture for **LOcation-FRee (LOFR)** HPE as in Fig. 2. Firstly, thanks to the self-attention mechanism in Transformer [40], the global contextual information can be effectively captured in HPE. To better capture the precise spatial information, we propose a **Multi-scale Spatial-guided Context Encoder (MSC-En)**. In MSC-En, we design a **Spatial-aware Position Encoding (SPE)** module helps the model focus on body regions while capturing the global human context. We capture the multi-scale feature representations to conduct self-attention learning leading to more comprehensive context from aggregated multi-scale information, which is robust to background clutters. For alleviating the location confusions, we equip the model with the structural relations encoded via GCN and propose a **Relation-Guided Pose Decoder (RGP-De)**. In RGP-De, a **Relation-encoded Pose Prototype Generation (RePPG)** module is designed to express keypoint-specific relations to help infer the confused parts during decoding. Finally, part-aware response maps denoting the spatial distributions of specific keypoints (*e.g.*, ankle or head) are acquired by exploring the interactions between human context memory and prototypes. To prompt diversity and fine-grain, a **Part Diversity Constraint (PDC)** is devised to encourage lower correlation between part features and force them focus on their own parts.

In a nutshell, the contribution of this paper is three-fold:

- To our knowledge, we are the first to develop a location-free HPE only with image-level labels. The effectiveness is extensively validated on three datasets and the performance can even outperform the supervised one when few location labels are given.
- We employ a multi-scale spatial-guided context encoder to capture the global context features and meanwhile make it attend to the body parts by the aid of the spatial-aware positional encoding.
- We design a relation-encoded pose prototype generation strategy to mine the inherent spatial relation prior between keypoints via GCN. Also, a part diversity constraint makes the part-aware features more distinguished.

## 2. Related Work

**Human Pose Estimation.** Recently, researchers have made painstaking efforts [9, 36, 45] to make progress on HPE, two mainstream methods are prevalent including bottom-up [18, 32] and top-down [6, 37, 42]. The former directly estimate all the keypoints and assign them into persons. The latter firstly detect the human bounding boxes and locate the keypoints within each box. However, the above works all solve the fully-supervised regression and no research explore the weakly-supervised HPE. This paper follows the top-down pipeline. After acquiring the bounding boxes, we acquire the keypoint locations from a perspective of classification with only the category labels.

**Transformers in computer vision.** Recently, Transformers attract much attention in computer vision. ViT [10] applies a pure Transformer framework to a series of image patches for classification. Besides, vision Transformer is widely applied to object detection [51], segmentation [52]. Further, DETR [5] and Deformable DETR [54] predict a box set for matching the object location.

Specially, the Transformer is also applied in HPE, including hand pose estimation [14] and 3D HPE [23, 50]. The most close to ours is the applications in 2D HPE [22, 24, 26, 44]. These studies achieve impressive performance implying that Transformer is suitable and effective for modeling human poses. Therefore, we also leverage the transformer to explore the weakly-supervised HPE with only the category labels.

**CAM-based WSOL.** Weakly-supervised object localization aims to locate the objects with only image-level labels. Since CAM was proposed in [53], CAM-based methods have been achieved great success for both WSOL and WSSS. For WSOL, CAM suffers from only identifying small discriminative region of objects. After that, a series of works [20, 46, 48] are proposed to improve the quality of CAM. However, these improvements are ineffective when extended to HPE.

In light of CAM, we rethink the HPE and aim to locate keypoints with the image-level category labels. Even if CAM-based methods have made success in WSOL, it doesn’t work well for the keypoint prediction. The local human parts and the subtle differences between categories undoubtedly bring great challenges to achieve accurate HPE.

## 3. Method

### 3.1. Framework Overview

As shown in Fig. 2, the LOFR framework primarily comprises of the MSC-En and RGP-De. Given an input, we firstly obtain the multi-scale feature representations through the CNN backbone. Then, the multi-level feature maps are processed with SPE to feed into the encoder to conduct the self-attention for capturing the human context memory  $F_m$ .

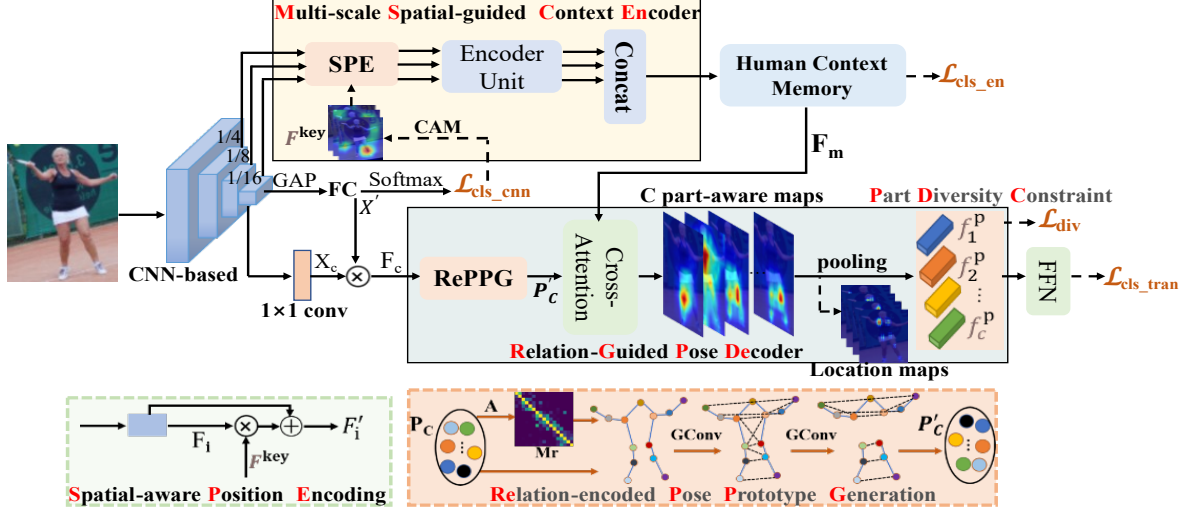


Figure 2. **LO**cation-**FR**ree (**LOFR**) HPE pipeline. It comprises a **MSC-En** with **SPE** to capture the multi-scale human context related to the part regions and a **RGP-De** integrated with **RePPG** that generate the encoded pose prototypes guiding the cross-attention with human context to parse more accurate part responses. The **PDC** helps learn more distinguished part features.

around body regions. In RGP-De, a set of pose prototypes  $P'_C$  that initialized with RePPG are sent into decoder performing cross-attention with the context memory decoding the part-aware response maps. Part features can be obtained by pooling operation where the location maps are treated as different spatial keypoint locations. Except the general binary cross entropy (BCE) loss for the predicted joint categories, a part diversity constraint is used to capture more distinguished part features.

### 3.2. CNN-based

Following the top-down pipeline, the detected single person images are fed into the CNN-based network to acquire the feature map  $X \in \mathbb{R}^{H \times W \times D}$  where  $H, W$ , and  $D$  are the height, width and channels. We turn  $X$  into  $X' \in \mathbb{R}^{H \times W \times C}$  with  $1 \times 1$  convolution and  $C$  is the number of keypoint classes. Then, followed by GAP, FC and a softmax layer for classification as in Fig. 2, we get the category-specific activation maps  $M = [m_1, m_2, \dots, m_c] \in \mathbb{R}^{H \times W \times C}$  by convolving the weights of FC with feature map  $X'$  as follows:

$$m_c(x, y) = \text{RELU}(\sum_k w_k^c x'_k(x, y)), \quad (1)$$

where  $w_k^c$  and  $x'_k$  depict the weight of the  $c$ -th category and  $k$ -th feature map, respectively.

For acquiring the initial node embedding for the graph in RGP-De (in sec 3.4), we compute category-specific keypoint vectors  $F_c = [f_1, f_2, \dots, f_c] \in \mathbb{R}^{C \times D'}$  as:

$$f_c = x'^T X^c = \sum_{i=1}^H \sum_{j=1}^W x'_{i,j} x^c_{i,j}, \quad (2)$$

where  $x'_{i,j}$ ,  $x^c_{i,j}$  are the  $(i, j)$ -th feature of feature maps and  $X^c \in \mathbb{R}^{H \times W \times D'}$  is got with a  $1 \times 1$  convolution.

### 3.3. Multi-scale Spatial-guided Context Encoder

We propose a MSC-En equipped with the SPE to capture the multi-scale spatial-aware human context information as shown in the yellow box of Fig. 2.

**Spatial-aware Position Encoding.** As discussed above, we have obtained the part-aware CAMs  $M$ . Further, we obtain coarse keypoint location maps  $\{F_i^{key}\}, i \in 1, 2, \dots, C$  via finding the max value of  $M$ . Rather than adopt the random initialized position embedding, we take  $\{F^{key}\}$  as the implicit position prior to help the model know spatial part locations while capturing the context. Given the input features  $F$ , we establish spatial-aware inputs as below:

$$\varphi = F \otimes F^{key}, \quad (3)$$

$$F' = \psi(\varphi \oplus F). \quad (4)$$

We view  $\varphi$  as the updated positional encoding to sum with  $F$ , the  $\psi$  depicts a feature transformation operation. The  $\otimes, \oplus$  depicts the cross-product and element-wise sum operation. The resulted  $F'$  is fed into the encoder.

**Multi-scale Context Learning.** For capturing more comprehensive context, we extract the multi-scale features  $\{F_i\} \in \mathbb{R}^{H \times W \times C}, i \in 1, 2, 3$  from the CNN backbone at the down-sampled ratios of  $1/4, 1/8, 1/16$ . We obtain their position encoding and the updated multi-scale input features  $\{F'_i\}, i \in 1, 2, 3$  in Eq. 3 to conduct self-attention (SA) mechanism.

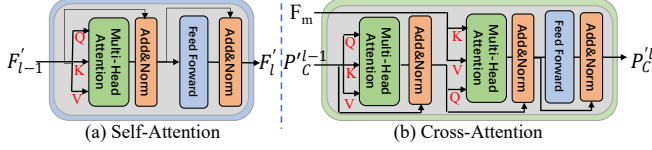


Figure 3. Illustration of the Encoder-Self-Attention and Decoder-Cross-Attention modules.

Given  $F'$  as input, the multiple SA modules are employed to learn pixel-level human contextual dependencies of multi-scale features. As depicted in Fig. 3 (a), a SA module consists of multi-head self-attention (MSA) and feed forward networks (FFN), layer normalization (LN) and residual connection is applied after every block. The FFN contains two linear layers with a ReLU. For  $l$ -th layer, the input to SA is a triplet of (query, key, value) computed from the input  $F'_{l-1}$  as:

$$Q = F'_{l-1}W_Q^l, K = F'_{l-1}W_K^l, V = F'_{l-1}W_V^l, \quad (5)$$

where  $W_Q^l \in \mathbb{R}^{C \times d_q}$ ,  $W_K^l \in \mathbb{R}^{C \times d_k}$ ,  $W_V^l \in \mathbb{R}^{C \times d_v}$  are the parameter matrices of linear projection heads, and  $d_q$ ,  $d_k$  and  $d_v$  is the dimensions of input. The SA is formulated as:

$$s_{qk} = \text{Softmax}\left(\frac{F'_{l-1}W_Q^l(F'_{l-1}W_K^l)^T}{\sqrt{d_k}}\right), \quad (6)$$

where the attention weights  $s_{qk}$  are calculated based on the dot-product similarity between each query and the keys. The  $d_k$  is a scaling factor modeling the inter-dependency between different spatial pixels of human part regions.

$$SA(F'_{l-1}) = F'_{l-1} + F'_{l-1}W_V^l s_{qk}, \quad (7)$$

the weighted sum of the values can aggregate these semantically related spatial pixels to update the context. Since pixels belonging to the human body part have high similarities while are distinct from background pixels, the feature captures more complete human body would be more robust to backgrounds. The MSA is an extension with  $h$  SA and projects their concatenated output as:

$$MSA(F'_{l-1}) = \text{Concat}(SA_1, \dots, SA_h)W_O^l, \quad (8)$$

where  $W_O^l \in \mathbb{R}^{hd_k \times C}$  is a parameter of linear head. We set  $h = 8$ ,  $C = 256$  and  $d_q, d_k, d_v$  equal to  $C/h = 32$ . Then, we use FFN to produce the context-aware memory:

$$F'_l = MSA(F'_{l-1}) + FFN(MSA(F'_{l-1})) \in \mathbb{R}^{L \times C}. \quad (9)$$

Up to now, the single-scale human context features are captured, which are robust to background clutters. For aggregating the multi-scale context, we concat the multi-level output  $\{F'_i\}$ ,  $i \in \{1, 2, 3\}$  as the final context memory  $F_m$ .

### 3.4. Relation-Guided Pose Decoder

In RGP-De, we involve a set of relation-encoded pose prototypes that contain the category-specific semantics and structural relations among joints to let the decoder decode more accurate part-aware response maps. An additional part diversity constraint enables the part locations more accurate and focused.

**Relation-encoded Pose Prototype Generation.** Notably, the human poses have inherent structure conforming to the kinematic constraint. For example, the adjacent or symmetrical joints are more likely to possess highly consistent semantic information. Therefore, we design RePPG to integrate joint-wise relations into the updated pose prototypes for parsing more precise part-aware locations.

We firstly introduce a set of pose prototypes  $\mathcal{P}_C = \{p_i\}_{i=1}^C$ ,  $p_i \in \mathbb{R}^{1 \times d}$  which determines whether pixels of the feature  $F_m$  belong to the part  $i$ . We initialize the category-specific keypoint features  $F_c$  in Eq. 2 as the node feature of  $\mathcal{P}_C$ . We build an intuitive graph  $G = (V, E)$  based on  $F_c$ .  $V = \{v_i, i = 1, 2, \dots, C\}$  is the node set to depict keypoints.  $E = \{v_i v_j \mid \text{if } i \text{ and } j \text{ are connected in the human body}\}$  refers to limbs of part. The adjacent matrix  $A = \{a_{ij}\}$  is initialized according to the pre-defined kinematic connection with  $a_{ij} = 1$  when  $v_i$  and  $v_j$  are adjacent in  $G$  or  $i = j$ , otherwise  $a_{ij} = 0$ .

Considering the human body structure is a natural graph with spatial constraint among joints, we thus model the keypoint relations via the recent SemGCN [49] to explore their structural relation. For a graph convolution, propagating features through neighbors helps to learn robust structure and captures the joint dependency. This equips pose prototypes with the structural relation prior, which is vital for activating the uncertain part responses while decoding.

Concretely, the updated node features are firstly gathered to node  $i$  from its neighbors  $j$ . The initial node features are collected into  $F_c^{(l)} \in \mathbb{R}^{D_l \times C}$  as in Eq. 2.

$$F_c^{(l+1)} = \vec{W}^{(l+1)} F_c^{(l)} \varphi_i(M_r \odot A), \quad (10)$$

where  $F_c^{(l)}$  and  $F_c^{(l+1)}$  are the node features before and after  $l$ -th conv.,  $\varphi_i$  is normalization,  $\vec{W}^{(l+1)}$  is the weight matrix. The  $M_r \in \mathbb{R}^{C \times C}$  denotes the local semantic relations between joints and updates with node features. In this way, the updated pose prototypes  $\mathcal{P}'_C = \{p'_i\}_{i=1}^C$  encoding the local semantic and spatial relation are obtained.

The cross-attention layer aims to learn more specific part-aware features through the interaction between  $F_m$  with the prototypes  $\mathcal{P}'_C$ . As shown in Fig. 3 (b), given  $F_m$ , queries come from prototypes  $\{p'_i\}_{i=1}^C$ , keys and values arise from the feature  $F_m$ . The implementation is the same as the above SA learning, the attention weights of all  $hw$  positions form a part-aware response map  $R_i = [r_{i,1}, r_{i,2}, r_{i,3}, \dots, r_{i,hw}]$ , which has high response values at



context features belonging to  $i$ -th part. We then obtain  $i$ -th part feature by pooling operation. By computing over all prototypes, we obtain  $C$  part response maps (each map is an attention map) and the corresponding part features  $\{F_i^p\}_i^C$ . Finally, we obtain the keypoint location map via finding the max value of their response map.

**Part Diversity Constraint.** A simple classification loss is insufficient to capture subtle differences among keypoint categories. Different prototypes may tend to focus on the same part (e.g., the main body), which may result in confusion among keypoint locations. Thus, making part features more distinguished, we design a part diversity constraint on  $\{F_i^p\}_i^C$  to let features attend to the corresponding local part.

$$\mathcal{L}_{div} = \frac{1}{C(C-1)} \sum_{i=1}^C \sum_{j=1, i \neq j}^C \frac{\langle f_i^p, f_j^p \rangle}{\|f_i^p\|_2 \cdot \|f_j^p\|_2}, \quad (11)$$

if the  $i$ -th and  $j$ -th part feature give a high weight to the same location, the  $\mathcal{L}_{div}$  will become large and prompt each part feature to adjust themselves adaptively.

### 3.5. Optimization

The classification output denotes joint-wise one-hots  $\mathcal{O}$ . Based on this, the overall objective function  $\mathcal{L}_{weak}$  includes three classification losses and a diversity loss. The classification loss  $\mathcal{L}_{cls}$  refers to the BCE, followed by CNN output, the encoder output and the final prediction, respectively.

$$\mathcal{L}_{cls} = \frac{1}{C} \sum_{i=c}^C BCE(\mathcal{O}_i, \mathcal{O}_i^*). \quad (12)$$

$$\mathcal{L}_{weak} = \alpha \mathcal{L}_{cls_{cnn}} + \alpha_1 \mathcal{L}_{cls_{en}} + \alpha_2 \mathcal{L}_{cls_{tran}} + \beta \mathcal{L}_{div}. \quad (13)$$

When we provide few location-labeled data, we adopt  $\mathcal{L}_{mse}^{sup}$  to measure the groundtruth  $H^*$  and the predicted heatmap  $H$ . The overall loss  $\mathcal{L}_{semi-weak}$  is depicted as,

$$\mathcal{L}_{mse}^{sup} = \frac{1}{C} \sum MSE(H^*, H), \quad (14)$$

$$\mathcal{L}_{semi-weak} = \mathcal{L}_{mse}^{sup} + \mathcal{L}_{weak}, \quad (15)$$

where  $\alpha$ ,  $\alpha_1$ ,  $\alpha_2$  and  $\beta$  are the weight factors, respectively.

## 4. Experiment

### 4.1. Datasets and Evaluation metric

**COCO Keypoint Detection** [25] consists of 118K training images, 20K testing images and the 5K validation images. The performance is evaluated by the OKS-based average precision (AP) and average recall (AR).

**MPII Human Pose Dataset** consists of 25K images with 40K objects, where 12K objects are for testing and the remaining for training. We use the standard PCKh [1] (head-normalized probability of correct keypoint) for evaluation.

**CrowdPose** contains 20K images and 80K human instances in three crowding levels by Crowd Index: easy ( $0 \sim 0.1$ ), medium ( $0.1 \sim 0.8$ ) and hard ( $0.8 \sim 1$ ). It aims to promote performance in crowded cases and adopts the same evaluation metrics as in MS-COCO.

### 4.2. Implementation Details

**Network Architectures.** Unless specified, the backbone adopts ResNet-50 and HR-w32. We choose DETR [5] as the Transformer baseline.

**Training.** We implement all experiments in PyTorch [33] on 4 Tesla V100s with 32GB. For MS-COCO, human detection boxes are resized to  $256 \times 192$  or  $384 \times 288$ . We adopt Adam [17] optimizer with a learning rate of  $4 \times 10^{-3}$  and a weight decay of  $10^{-4}$ . The learning rate of transformer is decreased by a factor of 10. For MPII, the input size adopts  $256 \times 256$  and  $384 \times 384$  and half-body augmentations are adopted. The training lasts for 160 epochs. For CrowdPose, the setting is similar with COCO and trained for 210 epochs. For data augmentation, we apply random flip and random resize with scale  $\in [0.65, 1.35]$  (cutout not used). The weight factors for the  $\alpha$ ,  $\alpha_1$ ,  $\alpha_2$  and  $\beta$  are set as 0.2, 0.2, 0.5 and 0.1, respectively.

### 4.3. Comparison with State-of-the-arts

#### 4.3.1 Location-free Setting

**On MS-COCO.** The result comparisons on MS-COCO test-dev are in Tab. 1. It's noted that the baseline is implemented based on the CNN (Res-50) with the original transformer-based architecture. The accuracy has sharply dropped compared with the supervised result. By contrast, ours achieves over 20% improvement than the baseline across all cases. This proves our LOFR can obtain more accurate keypoint locations. Although a certain gap exists with the supervised one, we still achieve a competitive performance especially compared with the bottom-up methods. We also achieve stable improvements even with different backbones and input sizes, which also reflect the good generality of our method.

**On MPII.** As shown in Tab. 2, only with the category labels, the baseline solely achieves 35.3% PCKh score. By comparison, the accuracy of ours boosts to 61.8% with a large margin of 26.5%. Additionally, our LOFR achieves consistent improvements on all kinds of joints, though a certain gap exists compared with the fully-supervised methods. It's possibly in that MPII contains in-the-wild images with diverse pose interactions which span from householding to outdoor sports thus it brings great challenges to the model only with category-level labels.

Table 1. Performance comparisons on COCO *test-dev* 2017. The best result is highlighted in bold, and same for other tables.

Method	Back	size	AP	AP <sub>50</sub>	AP <sub>75</sub>	AP <sub>M</sub>	AP <sub>L</sub>
<b>Bottom-up methods</b>							
G-RMI [32]	R101	353×257	64.9	85.5	71.3	62.3	70.0
AE [30]	-	512×512	65.5	86.8	72.3	60.6	72.6
PifPaf [19]	-	-	67.4	-	-	-	-
HigherNet [7]	HR32	512	66.4	87.5	72.8	61.2	74.2
HGG [15]	-	512	68.3	86.7	75.8	-	-
FCPose [27]	R101	800	65.6	87.9	72.6	62.1	72.3
DEKR [11]	HR32	512	70.7	87.7	77.1	66.2	77.8
<b>Top-down methods</b>							
CPN [6]	Incep	384×288	73.0	91.7	80.9	69.5	78.1
SBN [42]	R152	384×288	73.7	91.9	81.1	70.3	80.0
HRNet [37]	HR32	384×288	74.9	92.5	82.8	71.3	80.9
PoseFix [28]	R152	384×288	<b>76.7</b>	<b>92.6</b>	<b>84.1</b>	<b>73.1</b>	<b>82.6</b>
UDP [13]	R152	384×288	74.7	91.8	82.1	71.5	80.8
<b>Transformer-based methods</b>							
PRTR [22]	HR32	384×288	71.7	90.6	79.6	67.6	78.4
TFPose [26]	R50	384×288	72.2	90.9	80.1	69.1	78.8
TokenP [24]	HR32	256×192	74.7	89.8	81.4	71.3	81.4
TransP [44]	HR32	256×192	73.4	91.6	81.1	70.1	79.3
<b>Only w/ Category Labels</b>							
Baseline	R50	256×192	34.0	41.4	36.2	31.6	36.2
Baseline	R50	384×288	35.1	42.1	37.2	32.6	37.3
<b>Ours-LOFR</b>	R50	256×192	54.3	61.0	55.6	52.6	56.7
<b>Ours-LOFR</b>	R50	384×288	55.4	62.1	56.7	53.7	57.9
<b>Ours-LOFR</b>	HR32	256×192	54.8	61.8	56.1	53.2	57.3
<b>Ours-LOFR</b>	HR32	384×288	55.9	62.9	55.3	54.4	58.4
<b>Ours-LOFR</b>	HR48	256×192	55.5	62.4	56.7	53.8	57.8
<b>Ours-LOFR</b>	HR48	384×288	56.4	63.4	55.6	54.8	59.0

**On CrowdPose.** We further validate our method on the challenging CrowdPose dataset and the result is depicted in Tab. 3. The LOFR surpasses the baseline on all metrics, yielding the accuracy of mAP 42.5% with an improvement of 20.0%. Even for the AP (*hard*), we still bring a large improvement to 34.1%(+17.5%). This suggests our method is solid even toward the extreme crowded poses. The qualitative comparison results are shown in Fig. 4.

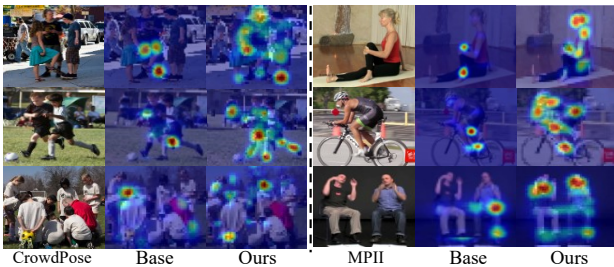


Figure 4. Qualitative comparisons on MPII and CrowdPose.

Table 2. Performance comparisons of PCKh@0.5 on MPII *test* set. \* means extra labels and larger image size are used.

Method	Hea	Sho	Elb	Wri	Hip	Kne	Ank	Total
Wei [41]	97.8	95.0	88.7	84.0	88.4	82.8	79.4	88.5
Newell [29]	98.2	96.3	91.2	87.2	89.8	87.4	83.6	90.9
Sun [36]	98.1	96.2	91.2	87.2	89.8	87.4	84.1	91.0
Tang [39]	97.4	96.4	92.1	87.7	90.2	87.7	84.3	91.2
Ning [31]	98.1	96.3	92.2	87.8	90.6	87.6	82.7	91.2
Chu [9]	98.5	96.3	91.9	88.1	90.6	88.0	85.0	91.5
Chou [8]	98.2	96.8	92.2	88.0	91.3	89.1	84.9	91.8
Yang [45]	98.5	96.7	92.5	88.7	91.1	88.6	86.0	92.0
Ke [16]	98.5	96.8	92.7	88.4	90.6	89.3	86.3	92.1
Xiao [42]	98.5	96.6	91.9	87.6	91.1	88.1	84.1	91.5
Tang [38]	98.4	96.9	92.6	88.7	91.8	89.4	86.2	92.3
Sun [37]	98.6	96.9	92.8	89.0	91.5	89.0	85.7	92.3
Su* [35]	98.7	97.5	94.3	90.7	93.4	92.2	88.4	93.9
Bin* [2]	<b>98.9</b>	<b>97.6</b>	<b>94.6</b>	<b>91.2</b>	93.1	<b>92.7</b>	89.1	<b>94.1</b>
Bulat* [3]	98.8	97.5	94.4	<b>91.2</b>	<b>93.2</b>	92.2	<b>89.3</b>	<b>94.1</b>
<b>Only W/ Category Labels</b>								
Baseline	53.5	73.0	33.8	26.5	15.7	14.6	7.1	35.3
<b>Ours-LOFR</b>	86.9	79.8	65.6	54.1	47.7	46.2	32.1	61.8

Table 3. Performance comparisons on CrowdPose *test* set.

Method	AP	AP <sub>50</sub>	AP <sub>75</sub>	AP <sub>M</sub>	AP <sub>H</sub>
<b>Bottom-up methods</b>					
OpenPose [4]	-	-	-	48.7	32.3
HigherHRNet [7]	67.6	87.4	72.6	68.1	58.9
DEKR [11]	68.0	85.5	73.4	68.8	58.4
<b>Top-down methods</b>					
Mask-RCNN [12]	57.2	83.5	60.3	57.9	45.8
SBN [42]	60.8	84.2	71.5	61.2	51.2
AlphaPose [21]	66.0	84.2	71.5	66.3	57.4
HRNet [37]	<b>71.7</b>	<b>89.8</b>	<b>76.9</b>	<b>72.7</b>	<b>61.5</b>
Baseline	21.5	39.1	28.2	23.4	16.6
<b>Ours-LOFR</b>	42.5	58.7	48.6	43.8	34.1

#### 4.3.2 Weakly Semi-supervised Setting

While we mainly focus on the image-level learning, our method can also achieve better performance when few location-labeled data are given. In this time, the training process is not changed but the labeled samples conduct supervised loss with groundtruth. We choose 5%, 10%, 25% labeled instances from the training set of MS-COCO and MPII, the remaining samples only have image-level labels and train as the location-free setting.

We validate the experiments on the *val* of the above datasets in Tab. 4. “Sup-only” means trained only with the samples labeled with groundtruth. And we reimplement the Sup-only baseline and the compared methods combined with the transformer baseline framework for the fair com-

parison. It shows that our model achieves stable improvements than the base *Sup-only* in all cases with the category label assistance. Noted that when 25% location labeled instances are given, our model achieves comparable performance with the fully-supervised (*ALL*) model. Moreover, the proposed strategy can benefit more accurate estimation under the fully-supervised setting and it achieves the result of 71.8% and 89.3% on COCO and MPII, outperforming the base (*ALL*) model by 0.8%, 0.6%, respectively.

Table 4. Result comparisons on *validation* set of the COCO and MPII datasets for different ratios of location data.

Dataset	Method	Back	5%	10%	25%	ALL
COCO	Sup-only [42]	R50	50.3	54.8	60.8	71.0
	Sup-only [37]	HR32	53.8	58.9	64.6	74.9
	SemiPose [43]	R50	57.7	61.6	66.4	-
	<b>Ours-WS</b>	R50	60.9	64.8	<b>70.6</b>	<b>71.8(+0.8%)</b>
	<b>Ours-WS</b>	HR32	64.6	68.9	<b>74.0</b>	<b>75.3(+0.5%)</b>
MPII	Sup-only [42]	R50	64.0	69.5	77.5	88.7
	SemiPose [43]	R50	71.3	76.3	82.5	-
	<b>Ours-WS</b>	R50	74.6	79.8	<b>88.1</b>	<b>89.3(+0.6%)</b>

#### 4.4. Ablation Study

**Effectiveness of SPE, MS and PDC.** In Tab. 5, the model with SPE improves 3.4% mAP than baseline. This reveals that it's useful to let model focus on the local part context learning. And multi-scale (MS) strategy further improves 1.8% and the PDC also improves 2.0% based on RePPG.

For further showing its effectiveness intuitively, we visualize the location maps produced by the model with or without SPE and PDC in Fig. 5. We observe that with SPE, the keypoint location is more accurate and complete, which illustrating our SPE provides spatial-aware guidance compared with the random initialized position encoding. The PDC also helps model discover more explicit part regions than baseline. Additionally, the learned features by MSC-En can cover more part-specific context by the aid of SPE compared with the Tran-En in Fig. 7.

Table 5. Ablation study of each module on COCO *val2017*.

Model	Baseline	SPE	MS	RePPG	PDC	mAP	mAR
1	✓					39.1	44.5
2	✓	✓				42.5	47.8
3	✓		✓			40.9	46.7
4	✓			✓		44.9	50.4
5	✓				✓	41.1	46.5
6	✓	✓	✓			44.3	49.4
7	✓			✓	✓	46.9	52.6
8	✓	✓	✓	✓	✓	<b>54.9</b>	<b>60.6</b>

**Effectiveness of RePPG.** In Tab. 5, the result of mAP improves to 44.9%(+5.8%) by adding RePPG on the baseline.

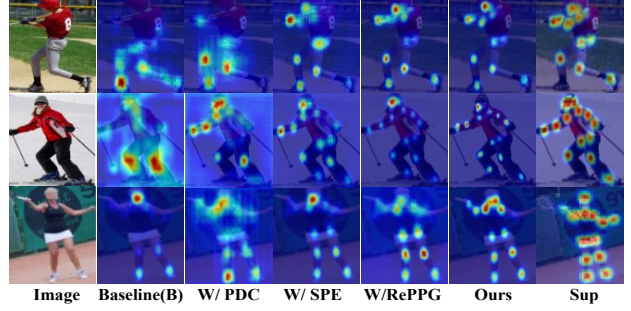


Figure 5. The ablated qualitative results for each module.

This powerfully indicates the necessity of mining the potential structural keypoint relations for guidance.

In our setting, the pose prototypes learn the statistical relevance between keypoints from the dataset, serving as the prior knowledge. To indicate the information encoded in these prototypes, we calculate their inner product matrix and visualize it in different cases in Fig. 6. Fig. 6 (d) reveals that one tends to be highly related to its symmetric or adjacent keypoints. For instance, the left hip is mostly related to right hip and left shoulder with high relevance score. Such finding conforms to our common sense and reveals what the model learns. But the Fig. 6 (b) mainly implies the self-correlation learning, which indicates that RePPG can encode more explicit local keypoint relations.

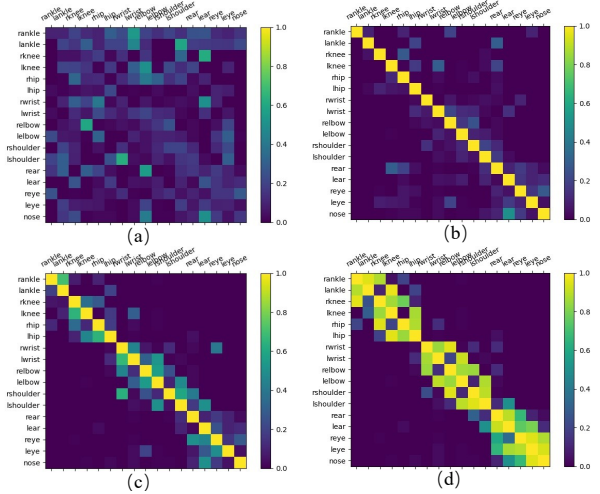


Figure 6. The inner product matrix of the learned pose prototypes. Each row of the matrix denotes the learned prior relations for a given type of keypoint with others. (a) depicts the random initialization; (b) depicts processed by SA; (c) the initial joint relevance; (d) depicts learned by RePPG.

**Impact of the model scaling.** In Tab. 6, we explore the effect of numbers of encoders and decoders in transformer. The performance grows at the first four layers and saturates as the layer increases and we choose the best setting.



Table 6. Effect of encoder ( $D_{En}$ ) and decoder ( $D_{De}$ ) numbers on COCO *val2017*.

Model	$D_{En}$	$D_{De}$	mAP	mAR
1	2	2	50.3	55.2
2	3	3	52.3	57.8
3	4	4	53.7	58.7
4	4	6	<b>54.9</b>	<b>60.6</b>
5	6	4	54.4	59.6
6	6	6	53.2	58.4

**Visualization and Analysis.** For illustrating our strategy intuitively, we visualize the detailed process in heatmap-based results in Fig. 7. Our RGP-De decodes more fine-grained and distinguished keypoint feature responses than the baseline. Our final result acquires more accurate locations even compared with the supervised result as in Fig. 5. Besides, we also visualize the qualitative comparison results in Fig. 8 and it also can prove the effectiveness of our method for the weakly-supervised HPE.

For ensuring the results are not cherry-picked, we compute the statistical average response value across joints on the whole dataset in Tab. 7. The joint response value of LOFR is obviously higher than the baseline. This illustrates that our model can acquire more explicit joint locations by the aid of the learned local context and guided relations.

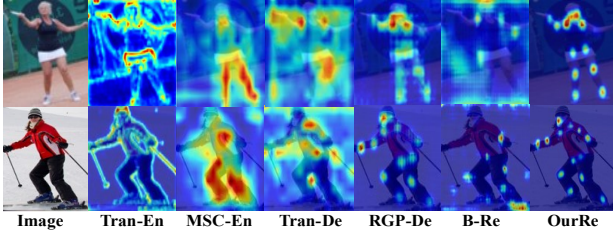


Figure 7. The original transformer-based encoder (Tran-En) vs. the proposed encoder (MSC-En) and the original decoder (Tran-De) vs. the proposed decoder (RGP-De). B-Re and OurRe depict the result of baseline and ours, respectively.

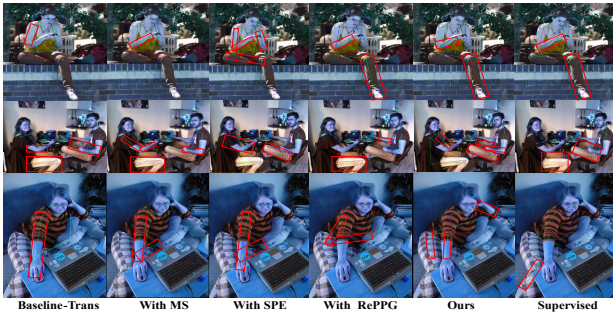


Figure 8. The qualitative results on random selected samples. We highlight the notable differences by a red rectangle box.

Table 7. Comparisons of the average response around the body joints across datasets. We normalize the value belongs to  $[0 \sim 1]$ .

Dataset	Ear	Eye	Nose	Hea	Sho	Elb	Wri	Hip	Kne	Ank
<b>Baseline</b>										
COCO	0.21	0.17	0.32	0.65	0.36	0.31	0.21	0.39	0.21	0.13
MPII	-	-	-	0.62	0.52	0.34	0.24	0.17	0.16	0.09
Crowd	-	-	-	0.61	0.32	0.28	0.16	0.35	0.15	0.11
<b>Ours-LOFR</b>										
COCO	0.45	0.38	0.52	0.86	0.71	0.53	0.46	0.61	0.54	0.42
MPII	-	-	-	0.85	0.76	0.64	0.53	0.48	0.50	0.35
Crowd	-	-	-	0.82	0.62	0.51	0.42	0.51	0.42	0.35

**Discussions.** Benefiting from the global feature learning capability of transformer, we can obtain more discriminative and comprehensive human body context than CNN. However, the body parts are very small and hard to distinguish, it’s vital to enable the transformer possess the position-aware prior to know where should be attended. Thus, we design the spatial-aware position encoding to focus locally. More significantly, the structural relations encoded via GCN can guide the decoder to activate the delicate part locations. With their collaboration, we achieve a relatively competitive location-free HPE.

Considering the issues of the image-label similarity, when a person appears completely, it possesses all keypoint categories. However, for the dataset we adopt, *e.g.*, COCO, according to the statistics in [34], the complete instances account for less than 50%, most instances only have half or a small number of body parts appearing in images. Therefore, the category labels across instances have sufficient diversity to enable model capture the distinct joint information. Noted that the cutout augmentation can bring about 0.8% improvement but it’s not involved in our method to avoid information leakage.

## 5. Conclusion and Limitation

In this paper, we shift the paradigm of human pose estimation from *Location-supervised* to *Location-free*. Accordingly, we propose a customized transformer-based HPE pipeline from the perspective of classification only with category-level labels. We firstly design a multi-scale spatial-guided context encoder to capture comprehensive context while focusing on the local part regions. To decode more accurate part-aware locations, we consider the inherent relation constraint among joints via GCN to encode the relational guidance into the pose prototypes. A part diversity constraint is used to keep the part features distinguished.

**Limitation.** The complexity of the model can be further reduced. And the model cannot better solve the occlusions for the multi-person estimation especially in crowded scenes (*e.g.*, on CrowdPose) and it worth exploring in future.



## References

- [1] Mykhaylo Andriluka, Leonid Pishchulin, Peter V Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. *CVPR*, pages 3686–3693, 2014. [5](#)
- [2] Yanrui Bin, Xuan Cao, Xinya Chen, Yanhao Ge, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, Changxin Gao, and Nong Sang. Adversarial semantic data augmentation for human pose estimation. *European Conference on Computer Vision*, pages 606–622, 2020. [6](#)
- [3] Adrian Bulat, Jean Kossaifi, Georgios Tzimiropoulos, and Maja Pantic. Toward fast and accurate human pose estimation via soft-gated skip connections. *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 8–15, 2020. [6](#)
- [4] Zhe Cao, Gines Hidalgo, Tomas Simon, Shiheng Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2019. [6](#)
- [5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pages 213–229. Springer, 2020. [2, 5](#)
- [6] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. *CVPR*, pages 7103–7112, 2018. [2, 6](#)
- [7] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5386–5395, 2020. [6](#)
- [8] Chiajung Chou, Juiting Chien, and Hwanntzong Chen. Self adversarial training for human pose estimation. *Asia Pacific Signal and Information Processing Association Annual Summit and Conference*, pages 17–30, 2018. [6](#)
- [9] Xiao Chu, Wei Yang, Wanli Ouyang, Cheng Ma, Alan L Yuille, and Xiaogang Wang. Multi-context attention for human pose estimation. *CVPR*, pages 5669–5678, 2017. [2, 6](#)
- [10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2020. [2](#)
- [11] Zigang Geng, Ke Sun, Bin Xiao, Zhaoxiang Zhang, and Jingdong Wang. Bottom-up human pose estimation via disentangled keypoint regression. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14676–14686, 2021. [6](#)
- [12] Kaifeng He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask r-cnn. *ICCV*, pages 2980–2988, 2017. [6](#)
- [13] Junjie Huang, Z. Zhu, F. Guo, and G. Huang. The devil is in the details: Delving into unbiased data processing for human pose estimation. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5699–5708, 2020. [6](#)
- [14] Lin Huang, Jianchao Tan, Jingjing Meng, Ji Liu, and Junsong Yuan. Hot-net: Non-autoregressive transformer for 3d hand-object pose estimation. In *ACM Multimedia*, pages 3136–3145, 2020. [2](#)
- [15] Sheng Jin, Wentao Liu, Enze Xie, Wenhao Wang, Chen Qian, Wanli Ouyang, and Ping Luo. Differentiable hierarchical graph grouping for multi-person pose estimation. *European Conference on Computer Vision*, pages 718–734, 2020. [6](#)
- [16] Lipeng Ke, Mingchun Chang, Honggang Qi, and Siwei Lyu. Multi-scale structure-aware network for human pose estimation. *ECCV*, pages 731–746, 2018. [6](#)
- [17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *ICLR*, 2015. [5](#)
- [18] Muhammed Kocabas and Emre Akbas. Multiposenet: Fast multi-person pose estimation using pose residual network. *ECCV*, pages 437–453, 2018. [2](#)
- [19] Sven Kreiss, Lorenzo Bertoni, and Alexandre Alahi. Pifpaf: Composite fields for human pose estimation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11977–11986, 2019. [6](#)
- [20] Krishna Kumar Singh and Yong Jae Lee. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In *ICCV*, pages 3524–3533, 2017. [1, 2](#)
- [21] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10863–10872, 2019. [6](#)
- [22] Ke Li, Shijie Wang, Xiang Zhang, Yifan Xu, Weijian Xu, and Zhuowen Tu. Pose recognition with cascade transformers. In *CVPR*, pages 1944–1953, 2021. [2, 6](#)
- [23] Wenhao Li, Hong Liu, Runwei Ding, Mengyuan Liu, and Pichao Wang. Lifting transformer for 3d human pose estimation in video. *arXiv preprint arXiv:2103.14304*, 2021. [2](#)
- [24] Yanjie Li, Shoukui Zhang, Zhicheng Wang, Sen Yang, Wankou Yang, Shu-Tao Xia, and Erjin Zhou. Tokenpose: Learning keypoint tokens for human pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11313–11322, 2021. [2, 6](#)
- [25] Tsungyi Lin, Michael Maire, Serge J Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and C Lawrence Zitnick. Microsoft coco: Common objects in context. *ECCV*, pages 740–755, 2014. [5](#)
- [26] Weian Mao, Yongtao Ge, Chunhua Shen, Zhi Tian, Xinlong Wang, and Zhibin Wang. Tfpote: Direct human pose estimation with transformers. *arXiv preprint arXiv:2103.15320*, 2021. [2, 6](#)
- [27] Weian Mao, Zhi Tian, Xinlong Wang, and Chunhua Shen. Fcpote: Fully convolutional multi-person pose estimation with dynamic instance-aware convolutions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9034–9043, 2021. [6](#)
- [28] Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Posefix: Model-agnostic general human pose refinement net-

- work. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7765–7773, 2019. 6
- [29] Alejandro Newell and Jia Deng. Stacked hourglass networks for human pose estimation. *ECCV*, pages 483–499, 2016. 6
- [30] Alejandro Newell, Zhiao Huang, and Jia Deng. Associative embedding: End-to-end learning for joint detection and grouping. *NIPS*, pages 2277–2287, 2017. 6
- [31] Guanghan Ning, Zhi Zhang, and Zhihai He. Knowledge-guided deep fractal neural networks for human pose estimation. *IEEE Transactions on Multimedia*, pages 1246–1259, 2018. 6
- [32] George Papandreou, Tyler Zhu, Nori Kanazawa, Alexander Toshev, Jonathan Tompson, Chris Bregler, and Kevin P Murphy. Towards accurate multi-person pose estimation in the wild. *CVPR*, pages 3711–3719, 2017. 2, 6
- [33] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. *NIPS*, 2017. 5
- [34] Matteo Ruggero Ronchi and Pietro Perona. Benchmarking and error diagnosis in multi-instance pose estimation. *Proceedings of the IEEE international conference on computer vision*, pages 369–378, 2017. 8
- [35] Zhihui Su, Ming Ye, Guohui Zhang, Lei Dai, and Jianda Sheng. Cascade feature aggregation for human pose estimation. *arXiv preprint arXiv:1902.07837*, 2019. 6
- [36] Ke Sun, Cuiling Lan, Junliang Xing, Wenjun Zeng, Dong Liu, and Jingdong Wang. Human pose estimation using global and local normalization. *ICCV*, pages 5600–5608, 2017. 2, 6
- [37] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. *CVPR*, pages 5693–5703, 2019. 2, 6, 7
- [38] Wei Tang, Pei Yu, and Ying Wu. Deeply learned compositional models for human pose estimation. *ECCV*, pages 197–214, 2018. 6
- [39] Zhiqiang Tang, Xi Peng, Shijie Geng, Lingfei Wu, Shaoting Zhang, and Dimitris N Metaxas. Quantized densely connected u-nets for efficient landmark localization. *ECCV*, pages 348–364, 2018. 6
- [40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017. 2
- [41] Shiheng Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. *CVPR*, pages 4724–4732, 2016. 6
- [42] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. *ECCV*, pages 472–487, 2018. 2, 6, 7
- [43] Rongchang Xie, Chunyu Wang, Wenjun Zeng, and Yizhou Wang. An empirical study of the collapsing problem in semi-supervised 2d human pose estimation. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11240–11249, 2021. 7
- [44] Sen Yang, Zhibin Quan, Mu Nie, and Wankou Yang. Transpose: Keypoint localization via transformer. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11802–11812, 2021. 2, 6
- [45] Wei Yang, Shuang Li, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. Learning feature pyramids for human pose estimation. *ICCV*, pages 1290–1299, 2017. 2, 6
- [46] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019. 1, 2
- [47] Fei Zhang, Chaochen Gu, Chenyue Zhang, and Yuchao Dai. Complementary patch for weakly supervised semantic segmentation. pages 7242–7251, 2021. 1
- [48] Xiaolin Zhang, Yunchao Wei, and Yi Yang. Inter-image communication for weakly supervised localization. *ECCV*, pages 271–287, 2020. 1, 2
- [49] Long Zhao, Xi Peng, Yu Tian, Mubbasir Kapadia, and Dimitris N Metaxas. Semantic graph convolutional networks for 3d human pose regression. *CVPR*, pages 3425–3435, 2019. 4
- [50] Ce Zheng, Sijie Zhu, Matias Mendieta, Taojiannan Yang, Chen Chen, and Zhengming Ding. 3d human pose estimation with spatial and temporal transformers. *arXiv preprint arXiv:2103.10455*, 2021. 2
- [51] Minghang Zheng, Peng Gao, Xiaogang Wang, Hongsheng Li, and Hao Dong. End-to-end object detection with adaptive clustering transformer. *arXiv preprint arXiv:2011.09315*, 2020. 2
- [52] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *CVPR*, pages 6881–6890, 2021. 2
- [53] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, pages 2921–2929, 2016. 1, 2
- [54] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *ICLR*, 2020. 2