LAMP: Leveraging Language Prompts for Multi-person Pose Estimation

Shengnan Hu, Ce Zheng, Zixiang Zhou, Chen Chen, and Gita Sukthankar

Abstract—Human-centric visual understanding is an important desideratum for effective human-robot interaction. In order to navigate crowded public places, social robots must be able to interpret the activity of the surrounding humans. This paper addresses one key aspect of human-centric visual understanding, multi-person pose estimation. Achieving good performance on multi-person pose estimation in crowded scenes is difficult due to the challenges of occluded joints and instance separation. In order to tackle these challenges and overcome the limitations of image features in representing invisible body parts, we propose a novel prompt-based pose inference strategy called LAMP (Language Assisted Multi-person Pose estimation). By utilizing the text representations generated by a well-trained language model (CLIP), LAMP can facilitate the understanding of poses on the instance and joint levels, and learn more robust visual representations that are less susceptible to occlusion. This paper demonstrates that language-supervised training boosts the performance of single-stage multi-person pose estimation, and both instance-level and joint-level prompts are valuable for training. The code is available at **https://github.com/shengnanh20/LAMP**.
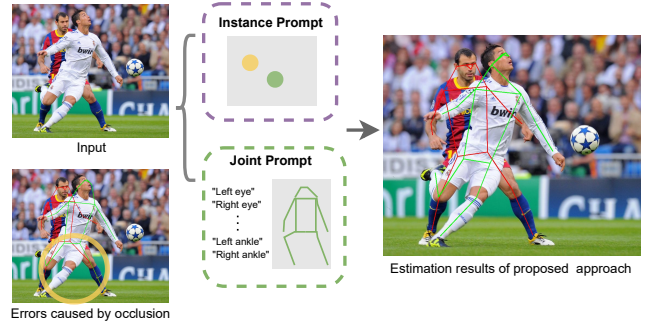
Fig. 1. Previous methods suffer from the difficulty of detecting overlapped persons and occluded keypoints. Our approach utilizes both instance and joint prompts. The image model is trained to maximize the correlation between text and image features. The results on the occluded scenario (right) demonstrate the superiority of **LAMP**.

I. INTRODUCTION

Multi-person pose estimation (MPPE) algorithms must simultaneously solve two problems: detecting the people in the scene and localizing the joint keypoints. Dealing with occlusion is an important challenge for human pose estimation, especially in scenes with multiple persons and complex contexts. While existing algorithms have achieved promising results with both top-down and bottom-up methods, they often fail to recover occluded people and joints [1]. Top-down methods [2], [3], [4], [5], [6], [7] usually rely on bounding-box cropping and cannot easily recover from instance detection failures or bounding-box overlap. Bottom-up methods [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19] experience difficulty estimating poses when there are overlapping individuals or occluded body parts, since they rely on grouping individual keypoints into poses. Also, bottom-up methods often require complex post-processing steps, which can be challenging to implement. Unlike top-down and bottom-up methods, the newer single-stage techniques [20], [21], [22], [23], [24], [25], [26] have a simpler architecture and training process, which can reduce the risk of error accumulation. Since they don't rely on grouping individual keypoints, single-stage methods can be more efficient and typically employ dense regression to localize body keypoints directly from scene feature vectors. Due to the difficulty of long-range location regression, improving the accuracy of the regression procedure remains a challenge.

With the recent improvements of multi-modal training, language-supervised visual representation learning has shown great success on multiple tasks, including segmentation [27], [28], depth estimation [29] and action recognition [30]. CLIP (Contrastive Language-Image Pretraining) [31] is one of the most popular pre-training models, and we leverage it for our research. Inspired by this line of research, this paper investigates the question: *can language supervised training improve multi-person pose estimation by improving the detection of occluded figures and body keypoints?* Our ultimate aim is to imbue a social robot with a human-centric visual understanding model that will enable it to navigate crowded streets, parks, and public places. Our MPPE model can be used as an input for downstream tasks such as human mesh recovery [32], [33], action recognition [34], and person-to-person social interactions [35].

Our proposed architecture (**LAMP**) (**L**anguage **A**ssisted **M**ulti-person **P**ose estimation) uses text prompts to link semantic descriptions with both instances of people in the scene and joint keypoints, transforming the MPPE problem into a set of contrast tasks. **LAMP** utilizes both instance prompts and joint prompts and trains the image model to maximize the correlation between text and image features. Prompts are generated automatically with templates and are only used during the supervised training phase. *Instance prompts* describe the relative locations of the people in the scene and whether or not they are occluded. This information is used to refine instance-aware heatmaps and helps **LAMP**
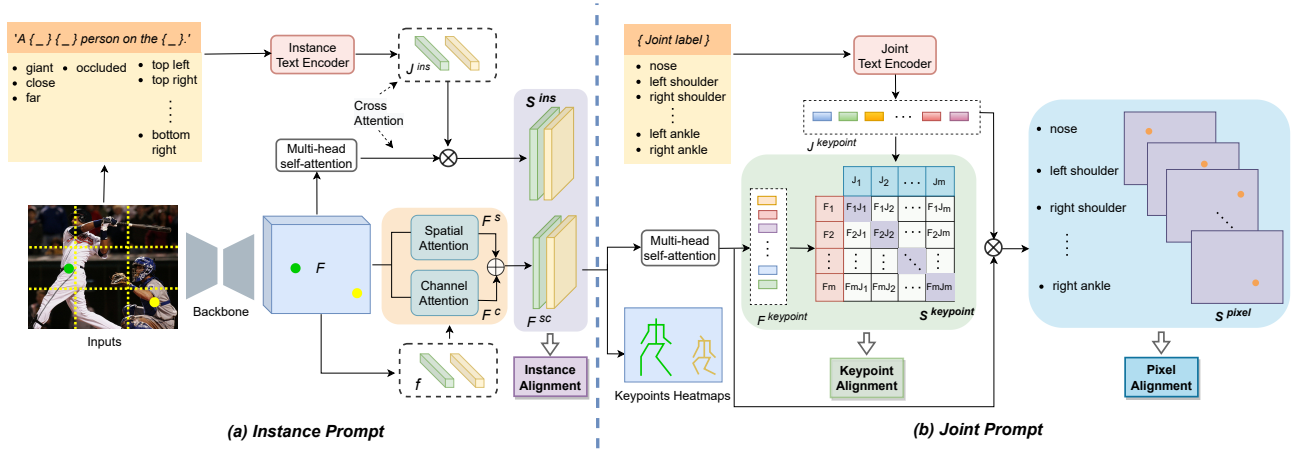
Fig. 2. Overview of the proposed architecture **LAMP**. (a) Given an input image, a backbone is applied to generate the global feature F. Then an image encoder with spatial attention and channel attention extracts a set of instance-aware features F^{sc} from the global features. A text encoder takes the description of each instance (Instance Prompt) as input and outputs a set of text features J^{ins}. A similarity-based textual embedding for each instance S^{ins} is then generated by computing the inner product between the instance textual feature and the global image feature. (b) The instance-aware features are utilized by the Joint Prompt module which takes the joint labels as input and generates a list of textual features. Then a matching evaluation between image-text models for the joints is conducted on two levels: keypoint level and pixel level. **LAMP** trains the image model to maximize the correlation between the textual embeddings and image features for the alignment modules shown in the figure.

ignore extraneous people in the scene. The instance-aware heatmaps are then used for joint localization. *Joint prompts* contain text describing the body keypoint (e.g. nose, left ankle, right shoulder). The CLIP model is trained on sufficient data to enable these joint descriptions to be semantically meaningful. To improve joint localization, we incorporate both keypoint-level alignment and pixel-level alignment. The intuition is that the inclusion of language pre-training makes MPPE more robust to occlusion, while also improving the performance of pixel-level alignment. Figure 2 shows the **LAMP** architecture.

Although CLIP is a tremendously valuable tool, adapting it to perform well on downstream applications such as multi-person pose estimation is not straightforward. This paper evaluates the performance of **LAMP** on OCHuman [36] and CrowdPose [37]. **LAMP** outperforms the other single-stage and bottom-up methods and is competitive against the best top-down methods. We also perform an ablative study to show the relative contributions of the instance and joint components of the vision-language model on **LAMP**'s performance.

This paper makes the following contributions:

- introduces a new architecture for multi-person pose estimation for social robots that incorporates language-assisted visual representation learning;
- demonstrates the power of language for making the image model robust to occlusion and improving the accuracy of the dense regression;
- outperforms existing single-stage models for MPPE on two challenging benchmarks for crowded pose estimation: OCHuman and CrowdPose.

II. RELATED WORK

A. Multi-person Pose Estimation

Multi-person pose estimation (MPPE) comprises two problems: 1) detecting all instances of people in the scene and 2) localizing their body keypoints. MPPE methods need to overcome both occlusion created by crowded scenes, as well as self-occlusion, where one body part obscures another. A secondary problem is handling scale and orientation diversity between people in the scene; it is difficult to perform human pose estimation on figures that may be only a few pixels in height.

1) Top-down: Top-down methods [2], [3], [4], [5], [6], [7] typically identify a bounding box for each person in the scene and then perform single person pose estimation. This two stage process often makes top down detectors computationally slower but gives them an edge on average precision. Mask-RCNN [2] reuses feature maps from the detector for pose estimation through the RoIalign operation, making it more efficient than other top down methods. HRNet [5] seeks to improve pose estimation performance by maintaining high resolution representations throughout the whole process, rather than recovering them from a high-to-low resolution network. We use the HRNet [5] as a backbone for several of the benchmarks. Xiao et al. [3] published a comprehensive evaluation of the effect of different detectors, backbone choices, and pose estimation algorithms on top down MPPE performance, and we benchmark **LAMP** vs. their simple baselines in addition to other SOTA algorithms.

2) Bottom-up: Bottom-up algorithms [8], [9], [10], [11], [12], [13], [14], [15], [16], [18], [17] group identity-free keypoints into persons using a variety of methods; for instance, DeepCut [18], [17] formulates the grouping problem as an

integer linear program. Other grouping methods include part linking [8], [11], associative embedding [9], and hierarchical clustering [10]. Separate heatmaps are usually calculated for each keypoint; a strength of these methods is that the heatmaps often outperform direct pixel-wise regression at localizing keypoints [38]. In general, bottom-up methods are more computationally efficient than top-down techniques due to their simpler convolutional pipelines but exhibit slightly inferior performance.

3) Single-stage: Single-stage approaches [20], [21], [22], [23], [24], [25], [26] densely regress candidate body keypoints directly from the original image, creating an end-to-end trainable pipeline. These methods are better at preserving valuable contextual information from the scene that gets discarded by both top-down and bottom-up systems. Also they do not rely on specific modules to perform RoI cropping, non-maximal suppression, and keypoint grouping. A key concern for these methods is improving the accuracy of the regression procedure. Contextual Instance Decoupling (CID) [26] improves the quality of regression in MPPE by separating the multi-person heatmap into instance-aware heatmaps that are later used to estimate heatmaps and keypoints for each person in the scene. We focus on further improving the performance of single-stage methods through the additional vision-language pretraining.

B. Vision-Language Models

Language has been shown to facilitate the learning of effective visual representations for zero-shot domain transfer on downstream tasks. CLIP (Contrastive Language-Image Pre-training) [31], a large-scale language-vision model, trains a textual encoder and a visual encoder on a dataset of 400M image-text pairs to explore the relationship between the modalities. Usually CLIP [31] is provided with a text prompt engineered for the downstream task to create a contrast task. The inclusion of the prompt during model training makes the downstream task more similar to the task originally used to create to model. Liu et al. [39] present a survey of prompt creation approaches. Although many of the downstream tasks that have benefited from CLIP [31] are simpler detection and classification tasks, DepthCLIP [29] illustrates that it is possible to leverage CLIP for pixel level tasks. PromptPose [40] uses vision-language pre-training for animal pose estimation; they also use prompt and pixel-level loss functions for body keypoint localization but do not learn instance level models.

III. METHOD

Our work leverages the text representations generated by language models such as CLIP [31] to facilitate the understanding of poses. Towards this end, we conduct a comprehensive study on the design and generation of pose prompts including instance level information and joint level information, which empowers the pose estimation baseline model in scenarios with multiple people.

A. Overview

Vision-language models have achieved recent successes in enhancing the performance of many vision tasks, which shows promising transfer capability between vision and text domains. Specifically, given a paired image and text $(image, text)$, vision-language models learn a multi-modal embedding space by jointly training an image encoder and text encoder to maximize the cosine similarity between the correct pair while minimizing the incorrect pairs [31]. Using a trained embedding, vision-language models can achieve zero-shot transfer to downstream image recognition tasks. Therefore, inspired by the previous work on introducing language supervision into visual recognition tasks [29], [40], we introduce a learned prior and investigate the use of text prompts to enhance the recognition of instances and keypoints in multi-person pose estimation.

Figure. 2 shows the flow of our instance-joint prompt pipeline for multi-person pose estimation. There are two key modules in the proposed architecture: instance prompt and joint prompt, where the instance prompt is used to provide dense instance-related information and semantics, and the joint prompt is used to refine the multi-modal connections at a joint level.

Given an image $I \in \mathbb{R}^{3 \times H \times W}$, we extract multi-scale image features from the backbone (i.e. HRNet [5]). The multi-scale features are then concatenated together and generate the global image feature map F. After that, we employ the instance decoupling module [26] with spatial attention and channel attention as the image encoder, to obtain the image features for each instance. An instance prompt with a pre-trained language-image encoder (i.e. CLIP) is simultaneously used to exploit the instance-related features in the textual model thus providing effective supervision to improve the discriminative power of the instance encoder.

Instance-aware feature maps are then fed to the heatmap module with a CNN layer, to obtain the probability distribution maps for the keypoints, i.e. keypoint heatmaps.

In order to leverage the semantic information from the language model for accurate keypoint estimation, we propose a method that incorporates both keypoint-level and pixel-level contrast. This approach establishes the connections between the image features and the embedding prompts, allowing for a more effective utilization of joint-related semantics in the language model.

With the alignment training by the proposed model, we can align the instance and joint features between the image model and language model, thus providing an effective prompt to guide the image module to learn a powerful visual representation in the MPPE task.

B. Instance Prompt with Attribute Injection

1) Image Encoder: Following previous studies of attention mechanism [41], [42], [26], [43], we decouple the cues of each instance from the original feature map based on spatial attention and channel attention. Specifically, given a feature map F, we first roughly extract the features of instances based on the ground truth locations of their center

points, obtaining a set of instance features $f \in \mathbb{R}^{N \times C^{ori}}$. After that, a spatial attention module is used to separate instances into different spatial locations, while a channel attention module is used to decouple instances into different channels of the feature map.

To obtain the spatial feature of each instance, we compute an instance-aware mask M^i based on the location of the i-th instance following [26]. Then we generate the spatial recalibration as:

$$F_i^s = F \cdot M_i. \quad (1)$$

In order to decouple the channel feature map of the i-th instance, we perform an element-wise manipulation between the instance feature f_i and the global feature as follows:

$$F_i^c = F \otimes f_i, \quad (2)$$

where f_i is generated from the original feature based on features at the coordinates of the instance's center point.

After obtaining the spatial recalibration and channel recalibration of each instance, we then conduct a fusion recalibration to bolster the discriminative ability of the model among different instances. Thus the fused instance-aware features are calculated as:

$$F^{sc} = Relu(Conv([F_i^s; F_i^c])) \in \mathbb{R}^{N \times C \times H \times W}. \quad (3)$$

2) Textual Label Generation: At the core of our approach is the idea of prompting person pose perception from supervision contained in the language. Inspired by the findings in [44], [45] that an expressive description can benefit the transfer performance of vision language models, here we design a prompt for instance separation by injecting essential attributes of instances into the model.

The detailed structure of the instance-level prompt is illustrated in Figure 2 (a). For each instance $instance_i$ in the original image, we construct the prompt by applying the following template:

$$\begin{aligned} Prompt_i = Template(\{s^{Attr_m}\}, instance_i), \\ Attr_m \in \{location, depth, occlusion\}. \end{aligned} \quad (4)$$

Specifically, to calculate the location information of the instance, we divide the original image into 3×3 patches and label the instance using the coordinates of the center point of each instance's bounding box. To enhance the discrimination between different instances, we introduce a pseudo depth label based on the scale of each instance's bounding box. This is achieved by computing the ratio of the width and height of the bounding box to the width and height of the image. Based on this ratio, we assign approximate depth labels, such as "far," "close," and "giant," to the different instances within the image. This allows for a rough categorization of depth information based on the relative size of the instances. Another attribute that can distinguish the relationships between instances further is occlusion. By counting the visible keypoints of each instance, we can decide whether an instance is occluded, which can also help establish the depth orders of instances. Finally,

prompts formed by instance attributes will pass through the textual encoder of CLIP [31] into latent space by:

$$P^{ins} = TextualEncoder(Prompt_i) \in \mathbb{R}^{N \times C_{emb}}. \quad (5)$$

A transformer decoder layer with a multi-head self-attention mechanism is then applied to enhance the projection of the textual feature to the image feature, and we have the textual prompts $J_i^{ins} \in \mathbb{R}^{1 \times C_{emb}}$ for the i-th instance.

3) Instance-level Alignment: To explore the connections between the image feature and text embedded feature of each instance, it is necessary to project both of them into a multi-modal embedding space. Here for the image feature, we conduct a multi-head self-attention layer with a feed-forward network (FFN) after the original feature to get the normalized image encoded feature $F^{img} \in \mathbb{R}^{N \times C \times H \times W}$.

We then calculate the inner product between the mapped image features and embedded textual features J^{ins} and get an instance-pixel matching score map of each instance:

$$S_i^{ins} = F^{img} \cdot J_i^{ins} \in \mathbb{R}^{H \times W}. \quad (6)$$

With such a similarity-based ensemble, we could fuse the prior knowledge learned by CLIP [31] and the image feature generated by the backbone and achieve a more informative representation for each instance. Thus by conducting an alignment between the encoded image features F^{sc} and the embedded textual features S^{ins}, **LAMP** provides extra supervision and boosts the discriminative power of the instance image encoders. Finally, an MSE loss between the textual embeddings and the image feature is calculated to maximize the correlation between them.

C. Joint Prompt

As shown in Figure 2 (b), after obtaining the instance-aware image features, the joint-level prompt is designed to further refine the pose representation at a joint level. Inspired by [40], we enhance the textual prompt by utilizing the embeddings from two aspects: keypoint level and pixel level.

1) Keypoint-level Alignment: As mentioned in section III-B.3, the image features of instances F^{sc} are first fed into a multi-head self-attention layer to map into the multi-modal embedding space, obtaining $F^{ins} \in \mathbb{R}^{N \times C \times H \times W}$.

Then to establish the connections between the text and local keypoint features in the image, a keypoint-level alignment is conducted by exploring the feature similarities in the embedding domain. Concretely, given m joints, the textual encoder takes the description word of each body joint (e.g. nose, left shoulder, right shoulder) as input and then generates the textual features of the joints $J^{keypoint} \in \mathbb{R}^{m \times C_{emb}}$, where m is the number of joints.

To align the textual joint features with the local feature of the joints in the image, we utilize the ground truth locations of the joints to sample features from the original instance image features, obtaining keypoint local features $F^{keypoint} \in \mathbb{R}^{m \times C_{emb}}$. Then we calculate the cosine similarities between each $\{image, text\}$ pair of the joints and obtain the similarity map as:

$$S_i^{keypoint} = F_i^{keypoint} \cdot (J^{keypoint})^T \in \mathbb{R}^{m \times m}. \quad (7)$$

We then perform contrastive learning on the keypoint local features and textual embedding features based on the contrastive loss.

2) Pixel-level Alignment: Furthermore, to establish a dense connection between the image features and the joint textual features in the embedding space, a pixel-level prompt is introduced to enhance the keypoints regression.

Following section III-C.1, we obtain the textual embedding features of the joints $J^{keypoint} \in \mathbb{R}^{m \times C_{emb}}$. Here, we apply an inner product between the instance image features F^{ins} and the joint textual features in the embedding space and get the text-pixel matching score of each pixel location by:

$$S_i^{pixel} = F_i^{ins} \cdot J^{keypoint} \in \mathbb{R}^{H \times W}. \quad (8)$$

This score map is supervised by the target heatmaps through the pixel-level contrastive loss.

D. Loss Functions

1) Image Loss: Similar to [26], we train the instance decoupling attention model with a contrastive loss to enhance the discriminative power of each instance. Given a set of instance features $f \in \mathbb{R}^{N \times C^{ori}}$, we obtain:

$$L_{con}^{f_i} = -log \frac{exp(\|f_i\|^2 / \tau)}{\sum_n^{j=1} exp(\|f_i\| \|f_j\| / \tau)}, \quad (9)$$

where n is the total count of instances in the image, and τ is a temperature parameter, which is set to 0.5 in experiments.

For heatmap regression, we compute the Focal Loss [46] between the predicted heatmaps and ground truth heatmaps generated based on the keypoints, obtaining:

$$FL(p, h^{gt}) = \frac{-1}{N} \sum_{c=1}^{C} \sum_{x=1}^{H} \sum_{y=1}^{W}$$
$$\begin{cases} (1 - p_{cxy})^\alpha log(p_{cxy}) & \text{if } h_{cxy} = 1 \\ (1 - h_{cxy})^\beta (h_{cxy})^\alpha log(1 - h_{cxy})) & \text{otherwise,} \end{cases}$$
$$(10)$$

where h denotes the ground truth heatmaps and p denotes the predicted heatmaps. Thus we have the overall heatmap loss as:

$$L_{FL}^f = \frac{1}{N} \sum_{i=1}^{N} FL(p_i, h_i^{gt}). \quad (11)$$

2) Prompt Loss: After obtaining the instance textual embeddings, an MSE loss is computed between the textual embeddings and the image features of each instance, i.e.,

$$L_p^{ins} = MSE(S^{ins}, F^{sc}) \quad (12)$$

As for keypoint-level alignment, we perform a contrastive learning by computing the cross entropy classification loss between the obtained similarity map $S^{keypoint} \in \mathbb{R}^{m \times m}$ and the matching target $M_{gt} = [0, 1, 2..., m-1]$ as:

$$L_p^{keypoint} = \frac{1}{2}(CE(S^{keypoint}, M_{gt}) + CE((S^{keypoint})^T, M_{gt}) \quad (13)$$

To enable effective joint-pixel embedding, we utilize the ground truth heatmaps as the supervision of the joint-pixel matching map by computing the MSE loss as:

$$L_p^{pixel} = MSE(S^{pixel}, h^{gt}) \quad (14)$$

Formally, the overall loss function of our approach can be formulated as:

$$L = \lambda_1 L_{con}^f + \lambda_2 L_{FL}^f + \lambda_3 L_p^{ins} + \lambda_4 L_p^{keypoint} + \lambda_5 L_p^{pixel}, \quad (15)$$

where λ_{1-5} are the loss weights, respectively.

IV. EXPERIMENTS

A. Datasets and Metrics

We follow the protocol of the recent state-of-the-art multi-person pose estimation method (i.e. CID [26]) and evaluate the performance of the proposed method on two multi-person pose estimation benchmarks: OCHuman [36] and CrowdPose [37].

a) OCHuman: This is a challenging dataset with many crowded scenes. It contains 4731 images with 8110 persons in total, including 2500 for training and validation, and 2331 for testing. We report our results following the setting of previous methods [26] [47].

b) CrowdPose: This is another dataset with highly occluded scenarios. The dataset contains $10K$ images and $80K$ annotated person instances with 14 keypoints for training, $2K$ images for validation and $8K$ images for testing.

c) Evaluation Metric: We follow the standard evaluation metric [1] and use the Object Keypoint Similarity (OKS) metrics for pose estimation. We report average precision and average recall scores with different thresholds and different object sizes: $AP, AP^{50}, AP^{75}, AP^M, AP^L, AR, AR^M$ and AR^L.

B. Implementation Details

LAMP was implemented using Pytorch [48] and MM-Pose [49].[1] Following previous work [13], [26], we use the HRNet-W32 [5] pretrained on ImageNet as the backbone for the image encoder in the pose estimation pipeline. For the language model, we use the text encoder of CLIP-ViT-Base [31] and initialize the model with the corresponding pre-trained weights.

During training, the input images are resized to 512×512, followed by random rotation, scale jitter, and translation. We use the Adam optimizer [50] and set the initial learning rate as 0.001. Note that text encoders are frozen during training to prevent the language knowledge from being distracted [40].

For testing, images are resized with the short side to 512 while maintaining the aspect ratio between height and width.
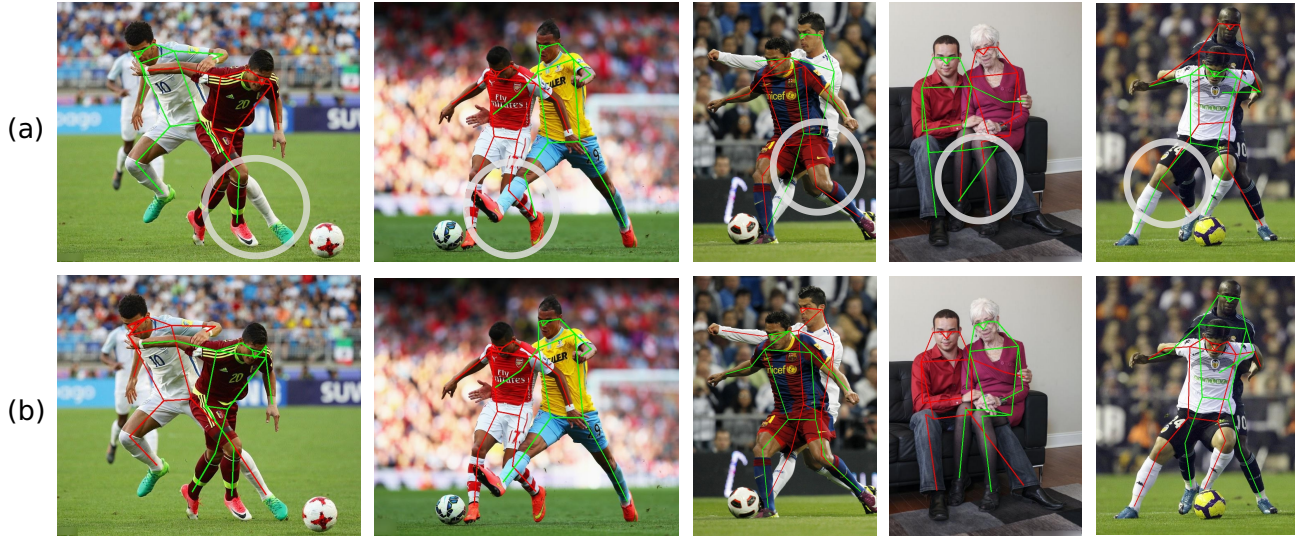
---

[1]Code will be released after publication.

Fig. 3. Qualitative comparison between the proposed method **LAMP** and the state-of-the-art approach CID on OCHuman testing set. (a) Estimation results produced by CID. (b) Estimation results produced by **LAMP**. Compared with CID, our proposed method **LAMP** is more robust to person overlapping and occlusion.

Method	$AP\uparrow$	$AP^{50}\uparrow$	$AP^{75}\uparrow$
Top-down methods			
Mask R-CNN [2]	20.2	33.2	24.5
SimplePose [3]	24.1	37.4	26.8
OPEC-Net [47]	32.8	60.5	31.1
Bottom-up methods			
HigherHRNet [12]	27.7	66.9	15.9
HGG [10]	36.0	-	-
Single-stage methods			
SPM [20]	47.6	67.5	53.2
DEKR (HRNet-W32) [13]	52.2	69.9	56.6
CID (HRNet-W32) [26]	57.5	75.5	63.3
Ours(HRNet-W32)	**58.8**	**76.6**	**64.3**

TABLE I

COMPARISONS WITH STATE-OF-THE-ART METHODS ON THE OCHUMAN TESTING SET.

C. Comparison with SOTA

a) Results on OCHuman Testing Set: To assess the performance of **LAMP** in tackling occlusions, we compare it with state-of-the-art methods on the OCHuman dataset, one of the most challenging benchmarks for crowded pose estimation. This dataset presents a significant challenge due to increased person overlapping and occlusions, which can negatively impact person detection and keypoint localization. As depicted in Table I, we compare three categories of methods: top-down (Mask R-CNN, SimplePose, and OPEC-Net), bottom-up (HigherHRNet and HGG), and single-stage (SPM, DEKR, and CID). Our results show that **LAMP** outperforms all top-down methods, demonstrating the superiority of our decoupling strategy in crowded scenes. It also outperforms bottom-up methods by achieving 22.8% AP higher than HGG. Additionally, **LAMP** surpasses single-stage methods,

with a 1.3% improvement over the most recent work CID.

b) Results on CrowdPose Testing Set: Table II presents the comparison of **LAMP** on another challenging benchmark CrowdPose. It can be observed that **LAMP** achieves the best performance, outperforming CID by 0.2%. A noteworthy finding is that the proposed model achieves a higher improvement over previous works on AP^{H}, outperforming CID by 0.4%. This further supports the efficacy of **LAMP** in handling heavy occlusions and overlapping people.

c) Qualitative Results: Fig. 3 visualizes some pose estimation results produced by the proposed model on the OCHuman testing set (bottom). This visual representation showcases the ability of our model to accurately estimate multiple poses in crowd scenarios, particularly for overlapped persons and occluded body parts.

D. Ablation Study

In this section, we perform an ablation study to investigate the contribution of each prompt component in the proposed model. The results are presented in Table III. We can observe that, by introducing the instance alignment and the keypoint alignment, the proposed approach improves the performance to 57.9% AP and 57.8% AP respectively. This validates the effectiveness of the instance prompt and joint prompt. Notably, the model obtains the largest performance improvement with the help of pixel alignment, and the complete **LAMP** further outperforms the baseline and achieves SOTA performance on the OCHuman dataset. This can be attributed to pixel-level alignment that maximizes the correlation between text prompts and image features more precisely, making the model more robust in the presence of overlapping persons or invisible joints. These results further demonstrate the superiority of the proposed **LAMP** architecture in occluded scenarios.

Method	Backbone	Input size	$AP\uparrow$	$AP^{50}\uparrow$	$AP^{75}\uparrow$	$AP^E\uparrow$	$AP^M\uparrow$	$AP^H\uparrow$
Top-down methods								
Mask-R-CNN [2]	ResNet-50	800	57.2	83.5	60.3	69.4	57.9	45.8
SimplePose [3]	ResNet-512	384x288	60.8	81.4	65.7	71.4	61.2	51.2
AlphaPose [51]	-		61.0	81.3	66.0	71.2	61.4	51.1
Bottom-up methods								
OpenPose [8]	VGG-19	-	-	-	-	62.7	58.7	32.3
HigherHRNet [12]	HRNet-W32	512	65.9	86.4	70.6	73.3	66.5	57.9
PINet [16]	HRNet-W32	512	68.9	88.7	74.7	75.4	69.6	61.5
Single-stage methods								
SPM [20]	Hourglass	-	63.7	85.9	68.7	70.3	64.5	55.7
DEKR(HRNet-W32) [13]	HRNet-W32	512	65.7	85.7	70.4	73.0	66.4	57.5
CID(HRNet-W32) [26]	HRNet-W32	512	71.2	89.8	76.7	77.9	71.9	63.8
E2Pose [52]	ResNet-512	512	66.5	-	-	-	-	-
Ours	HRNet-W32	512	**71.4**	**90.3**	**77.1**	**77.9**	**72.1**	**64.2**

TABLE II

COMPARISON WITH STATE-OF-THE-ART METHODS ON CROWDPOSE TESTING SET.

Model	$AP\uparrow$	$AP^{50}\uparrow$	$AP^{75}\uparrow$
Base (Image only)	57.5	75.5	63.3
Base + Instance Alignment	57.9	75.9	63.4
Base + Keypoint Alignment	57.8	76.2	63.8
Base + Pixel Alignment	58.6	76.5	63.7
Base + Instance + Keypoint + Pixel	**58.8**	**76.6**	**64.3**

TABLE III

ABLATION STUDY ON OCHUMAN.

V. CONCLUSION AND FUTURE WORK

Developing models of human-centered visual understanding that are robust to occlusion is an important stepping stone towards the development of social robots that can navigate crowded public places. In this paper, we tackle the challenge of multi-person pose estimation in crowded scenes. Our proposed **LAMP** architecture exhibits substantial improvements in occluded pose estimation through its end-to-end pipeline that leverages both instance and joint cues from the language model. Results from experiments on several crowded pose estimation benchmarks confirm the efficacy of our proposed method in addressing occlusion.

In future work, we will extend our model to handle downstream tasks such as detecting person-to-person social interactions and recognizing non-verbal communication. We envision our model as being useful for the human-robot interaction components of the Robocup@Home competition, assisting navigation and user control in UAV-human workspaces, and facilitating trust in human-robot teaming.

REFERENCES

[1] C. Zheng, W. Wu, C. Chen, T. Yang, S. Zhu, J. Shen, N. Kehtarnavaz, and M. Shah, "Deep learning-based human pose estimation: A survey," *ACM Comput. Surv.*, jun 2023. [Online]. Available: https://doi.org/10.1145/3603618

[2] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2961–2969.

[3] B. Xiao, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 466–481.

[4] K. Li, S. Wang, X. Zhang, Y. Xu, W. Xu, and Z. Tu, "Pose recognition with cascade transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1944–1953.

[5] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5693–5703.

[6] Y. Yuan, R. Fu, L. Huang, W. Lin, C. Zhang, X. Chen, and J. Wang, "Hrformer: High-resolution vision transformer for dense predict," *Advances in Neural Information Processing Systems*, vol. 34, pp. 7281–7293, 2021.

[7] Y. Ding, W. Deng, Y. Zheng, P. Liu, M. Wang, X. Cheng, J. Bao, D. Chen, and M. Zeng, "I2̂r-net: Intra-and inter-human relation network for multi-person pose estimation," *arXiv preprint arXiv:2206.10892*, 2022.

[8] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7291–7299.

[9] A. Newell, Z. Huang, and J. Deng, "Associative embedding: End-to-end learning for joint detection and grouping," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[10] S. Jin, W. Liu, E. Xie, W. Wang, C. Qian, W. Ouyang, and P. Luo, "Differentiable hierarchical graph grouping for multi-person pose estimation," in *European Conference on Computer Vision*. Springer, 2020, pp. 718–734.

[11] S. Kreiss, L. Bertoni, and A. Alahi, "Pifpaf: Composite fields for human pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11 977–11 986.

[12] B. Cheng, B. Xiao, J. Wang, H. Shi, T. S. Huang, and L. 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*, 2020, pp. 5386–5395.

[13] Z. Geng, K. Sun, B. Xiao, Z. Zhang, and J. Wang, "Bottom-up human pose estimation via disentangled keypoint regression," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14 676–14 686.

[14] G. Brasó, N. Kister, and L. Leal-Taixé, "The center of attention: Center-keypoint grouping via attention for multi-person pose estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 11 853–11 863.

[15] Z. Luo, Z. Wang, Y. Huang, L. Wang, T. Tan, and E. Zhou, "Rethinking the heatmap regression for bottom-up human pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 13 264–13 273.

[16] D. Wang, S. Zhang, and G. Hua, "Robust pose estimation in crowded scenes with direct pose-level inference," *Advances in Neural Information Processing Systems*, vol. 34, pp. 6278–6289, 2021.

[17] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele, "Deepercut: A deeper, stronger, and faster multi-person

pose estimation model," in *European Conference on Computer Vision (ECCV)*, 2016.

[18] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. Gehler, and B. Schiele, "Deepcut: Joint subset partition and labeling for multi person pose estimation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[19] N. Xue, T. Wu, G.-S. Xia, and L. Zhang, "Learning local-global contextual adaptation for multi-person pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 13 065–13 074.

[20] X. Nie, J. Feng, J. Zhang, and S. Yan, "Single-stage multi-person pose machines," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6951–6960.

[21] F. Wei, X. Sun, H. Li, J. Wang, and S. Lin, "Point-set anchors for object detection, instance segmentation and pose estimation," in *European Conference on Computer Vision*. Springer, 2020, pp. 527–544.

[22] W. Mao, Z. Tian, X. Wang, and C. Shen, "Fcpose: Fully convolutional multi-person pose estimation with dynamic instance-aware convolutions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9034–9043.

[23] D. Shi, X. Wei, X. Yu, W. Tan, Y. Ren, and S. Pu, "Inspose: instance-aware networks for single-stage multi-person pose estimation," in *Proceedings of the ACM International Conference on Multimedia*, 2021, pp. 3079–3087.

[24] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," *arXiv preprint arXiv:1904.07850*, 2019.

[25] D. Shi, X. Wei, L. Li, Y. Ren, and W. Tan, "End-to-end multi-person pose estimation with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 069–11 078.

[26] D. Wang and S. Zhang, "Contextual instance decoupling for robust multi-person pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 060–11 068.

[27] T. Lüddecke and A. S. Ecker, "Prompt-based multi-modal image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[28] J. Xu, S. De Mello, S. Liu, W. Byeon, T. Breuel, J. Kautz, and X. Wang, "Groupvit: Semantic segmentation emerges from text supervision," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[29] R. Zhang, Z. Zeng, Z. Guo, and Y. Li, "Can language understand depth?" in *Proceedings of the ACM International Conference on Multimedia*, 2022, pp. 6868–6874.

[30] C. Ju, T. Han, K. Zheng, Y. Zhang, and W. Xie, "Prompting visual-language models for efficient video understanding," in *European Conference on Computer Vision (ECCV)*. Springer, 2022.

[31] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8748–8763.

[32] C. Zheng, M. Mendieta, T. Yang, G.-J. Qi, and C. Chen, "Feater: An efficient network for human reconstruction via feature map-based transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[33] C. Zheng, X. Liu, G.-J. Qi, and C. Chen, "Potter: Pooling attention transformer for efficient human mesh recovery," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[34] Y. Hua, W. Wu, C. Zheng, A. Lu, M. Liu, C. Chen, and S. Wu, "Part aware contrastive learning for self-supervised action recognition," in *International Joint Conference on Artificial Intelligence*, 2023.

[35] K. Zhao, Y. Zhang, S. Wang, T. Beeler, and S. Tang, "Synthesizing diverse human motions in 3d indoor scenes," *arXiv preprint arXiv:2305.12411*, 2023.

[36] S.-H. Zhang, R. Li, X. Dong, P. Rosin, Z. Cai, X. Han, D. Yang, H. Huang, and S.-M. Hu, "Pose2seg: Detection free human instance segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 889–898.

[37] J. Li, C. Wang, H. Zhu, Y. Mao, H.-S. Fang, and C. Lu, "Crowdpose: Efficient crowded scenes pose estimation and a new benchmark," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 863–10 872.

[38] K. Sun, Z. Geng, D. Meng, B. Xiao, D. Liu, Z. Zhang, and J. Wang, "Bottom-up human pose estimation by ranking heatmap-guided adaptive keypoint estimates," *arXiv preprint arXiv:2006.15480*, 2020.

[39] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pretrain, prompt, and predict: A systematic survey of prompting methods in natural language processing," *CoRR*, vol. abs/2107.13586, 2021.

[40] X. Zhang, W. Wang, Z. Chen, J. Zhang, and D. Tao, "Promptpose: Language prompt helps animal pose estimation," *arXiv preprint arXiv:2206.11752*, 2022.

[41] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.

[42] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 3–19.

[43] Q. Zhao, C. Zheng, M. Liu, P. Wang, and C. Chen, "Poseformerv2: Exploring frequency domain for efficient and robust 3d human pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 8877–8886.

[44] S. Shen, C. Li, X. Hu, Y. Xie, J. Yang, P. Zhang, A. Rohrbach, Z. Gan, L. Wang, L. Yuan, *et al.*, "K-lite: Learning transferable visual models with external knowledge," *arXiv preprint arXiv:2204.09222*, 2022.

[45] Z. Qin, H. Yi, Q. Lao, and K. Li, "Medical image understanding with pretrained vision language models: A comprehensive study," *arXiv preprint arXiv:2209.15517*, 2022.

[46] H. Law and J. Deng, "Cornernet: Detecting objects as paired keypoints," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 734–750.

[47] L. Qiu, X. Zhang, Y. Li, G. Li, X. Wu, Z. Xiong, X. Han, and S. Cui, "Peeking into occluded joints: A novel framework for crowd pose estimation," in *European Conference on Computer Vision*. Springer, 2020, pp. 488–504.

[48] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.

[49] M. Contributors, "Openmmlab pose estimation toolbox and benchmark," https://github.com/open-mmlab/mmpose, 2020.

[50] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[51] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "Rmpe: Regional multi-person pose estimation," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2334–2343.

[52] M. Tobeta, Y. Sawada, Z. Zheng, S. Takamuku, and N. Natori, "E2pose: Fully convolutional networks for end-to-end multi-person pose estimation," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 532–537.