

# ProFD: Prompt-Guided Feature Disentangling for Occluded Person Re-Identification

Anonymous Authors

## 1 INTRODUCTION

The supplementary material offers extra details and additional experiments that were infeasible to include in the main article due to space constraints. The document is organized as:

- Details of Mask Generation. In this section, we will describe the mask generation process in detail.
- Additional Experimental Results. In this section, we will provide more ablation study and visualization results to fully demonstrate the effectiveness of **ProFD**.

## 2 DETAILS OF MASK GENERATION

In this paper, we define five regions based on the result of PifPaf[1]. Specifically, PifPaf can generate 17 part confidence and 19 affinity fields of input image. These part confidence and affinity fields can be regarded as probability maps presenting different body parts. Here, these 36 probability maps are manually divided into 5 groups as follows:

- **head**: "nose", "left eye", "right eye", "left ear", "right ear", "left eye to right eye", "nose to left eye", "nose to right eye", "left eye to left ear", "right eye to right ear", "left ear to left shoulder", "right ear to right shoulder".
- **upper arms and torso**: "left elbow", "right elbow", "left shoulder to left elbow", "right shoulder to right elbow", "left shoulder", "right shoulder", "left shoulder to right shoulder".
- **lower arms and torso**: "left wrist", "right wrist", "left elbow to left wrist", "right elbow to right wrist", "left hip", "right hip", "right shoulder to right hip".
- **legs**: "left hip", "right hip", "left knee", "right knee", "left ankle to left knee", "left knee to left hip", "right ankle to right knee", "right knee to right hip".
- **feet**: "left ankle", "right ankle".

Then, according to the aforementioned groups, the 5-parts probability map  $\mathcal{M} \in \mathbb{R}^{H' \times W' \times 5}$  are generated by performing pixel-wise max operation on these 36 probability maps. Finally, we perform a argmax operation on the probability map  $\mathcal{M}$  to generate body parsing label  $Y$ , as follows:

$$Y(h, w) = \begin{cases} 0 & \text{if } \max_c(\mathcal{M}(h, w, c)) < 0.5 \\ 1 + \arg \max_c(\mathcal{M}(h, w, c)) & \text{otherwise,} \end{cases} \quad (1)$$

where  $\mathcal{M}(h, w, c)$  denotes the probability of spatial location  $(h, w)$  belonging to group  $c$ .

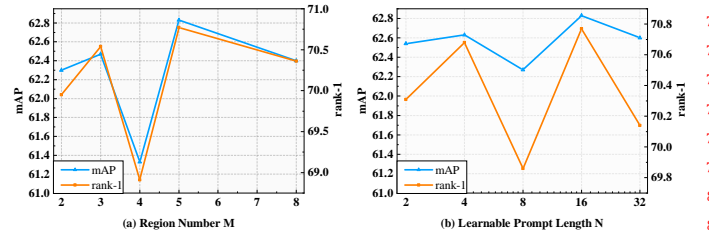
## 3 ADDITIONAL EXPERIMENTAL RESULTS

### 3.1 Ablation Study

**3.1.1 Effectiveness of Prompt Design.** To verify the impact of prompt design on model's performance, we designed three types of prompts: 'a photo of a {class}', 'a {class} part of a person', and '[Learnable Template] {class}'. The experimental results are shown in Table 1. We can

**Table 1: Performance of ProFD with different prompt designs on Occluded-Duke. {class} represents the class token, and [Learnable Template] represents learnable prompts.**

Prompts	Rank-1	mAP
"a photo of a {class}."	68.9	61.2
"a {class} part of a person."	69.3	61.5
"[Learnable Template] {class}"	<b>70.8</b>	<b>62.8</b>



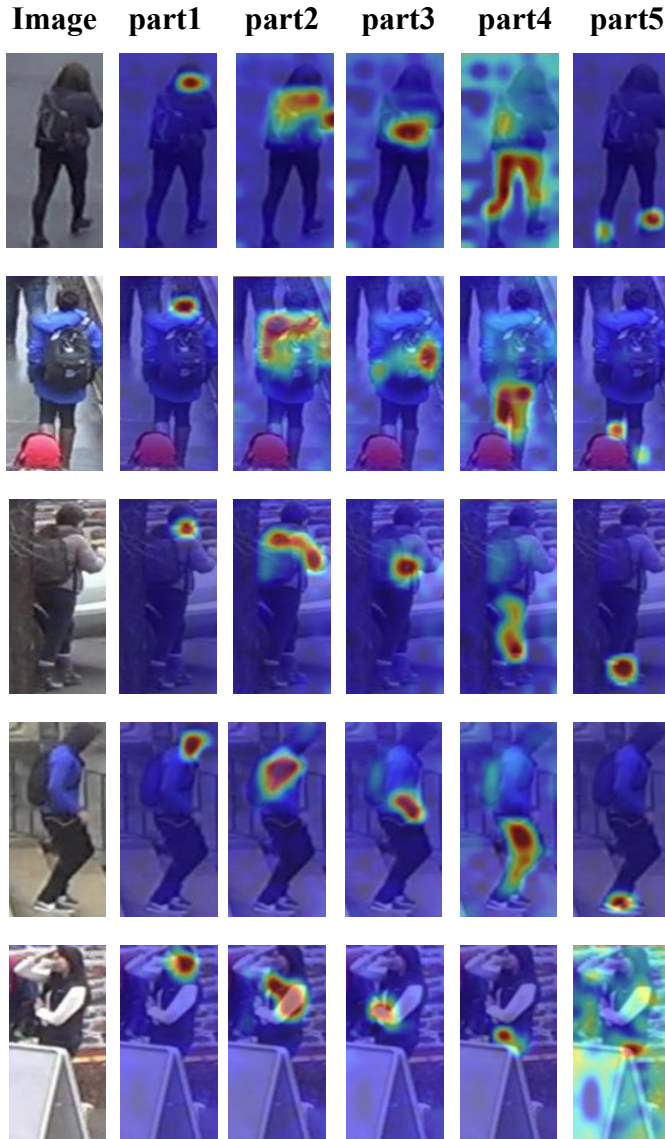
**Figure 1: Evaluation of the performance with different region number  $M$  and learnable prompt length  $N$  on Occluded-Duke.**

**Table 2: Five types of grouping strategy for mask generation.**

M	Grouping Strategy
2	upper body(head+arms+torso), lower body(legs+feet)
3	head, middle body(arms+torso), lower body(legs+feet)
4	head, middle body(arms+torso), legs, feet
5	head, upper arms and torso, lower arms and torso, legs, feet
8	head, left arm, right arm, torso, left leg, right leg, left foot, right foot

observe that using fixed prompt templates significantly degrades the model's performance. Compared to the manually designed templates, using a learnable template increases mAP by at least 1.3% and Rank-1 by at least 1.5%. However, it can also be observed that the second type of prompt performs better than the first type, as it provides more prior information to help the model better align visual-textual features to some extent.

**3.1.2 Evaluation of the Region Number.** To validate the influence of different region numbers  $M$ , we conducted numerous experiments by generating region masks using various grouping strategies, which are defined in Table 2. The experimental results are presented in Figure 1 (a). In general, we can observe that a bigger region number  $M$  leads to better performance. When the region number  $M$  is equal to 5, the best performance is achieved. However, performance begins to decline when the region number is greater than 8. This is because finer-grained partitioning of body parts may be more susceptible to noise in the mask itself, thereby affecting the final result. Here is a delicate balance: a smaller region number may be susceptible to occlusion effects, whereas a larger region number may be influenced by noisy masks from off-the-shelf model.



**Figure 2: Visualization of the hybrid attention. The hybrid attention, compared to the spatial-aware attention with noise, more accurately focuses on prompt-specific regions and also pays attention to some prominent features, such as a person’s backpack.**

**Table 3: Performance comparison of the occluded ReID problem on the Occluded-Duke, Occluded-ReID and P-DukeMTMC.**

Method	Occluded-Duke		Occluded-ReID		P-DukeMTMC	
	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
w/o Local mem	70.7	<b>62.9</b>	<b>91.2</b>	88.3	91.6	83.5
Concatenated mem	<b>70.8</b>	62.8	91.1	<b>88.5</b>	<b>91.7</b>	<b>83.7</b>
Separate mem	67.2	60.6	90.5	88.2	89.9	82.4

**3.1.3 Evaluation of the Length of Learnable Prompts.** To investigate the impact of learnable prompt lengths, we conducted experiments using five different lengths of learnable prompts: 2, 4, 8, 16, and 32. The experimental results are shown in Figure 1 (b). Overall, we can find that performance improves as the length  $N$  increases, reaching its peak when the length  $N$  equals 16. However, performance slightly declines when  $N$  is set to 32. This could be attributed to an excessive number of learnable prompts, which may negatively affect the model’s generalization ability, leading to catastrophic forgetting and over-fitting.

**3.1.4 Effectiveness of Self-distillation Strategy for Part Features.** As described in Section 4.3.2, their similarity of part features is independent of ID labels. And Due to the lack of annotations for part features, the way to distill knowledge for part features differs from global features. To thoroughly validate this conclusion, we conducted experiments on three occluded person ReID datasets, Occluded-Duke [2], Occluded-ReID [3], and P-DukeMTMC [4]. We compared three scenarios: training part features without self-distillation (Line 1), concatenating part features and training with a single memory bank (Line 2), and training with separate memory banks for each part feature (Line 3). The experimental results are presented in Table 3. We can observe that the third strategy significantly suppresses performance compared to the first two strategies. The decrease in performance occurs because utilizing global ID labels to distill part features results in the erosion of local similarities, consequently reducing the representativeness of these features.

## 3.2 Visualization

**Visualization of Hybrid Attention Map.** We present the visualization of hybrid attention map to in Figure 2. We can find that hybrid attention accurately focus to prompt-specific regions while also allocating attention to prominent features of person, aiding in the better identification. The results indicate that the hybrid attention has achieved its intended purpose.

## REFERENCES

- [1] Sven Kreiss, Lorenzo Bertoni, and Alexandre Alahi. Pifpaf: Composite fields for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11977–11986, 2019.
- [2] Jiaxu Miao, Yu Wu, Ping Liu, Yuhang Ding, and Yi Yang. Pose-guided feature alignment for occluded person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 542–551, 2019.
- [3] Jiaxuan Zhuo, Zeyu Chen, Jianhuang Lai, and Guangcong Wang. Occluded person re-identification. *ICME*, pages 1–6, 2018.
- [4] Zhimeng Zhang, Jianan Wu, Xuan Zhang, and Chi Zhang. Multi-target, multi-camera tracking by hierarchical clustering: Recent progress on dukemtmc project. *ArXiv*, abs/1712.09531, 2017.