Diverse Part Discovery: Occluded Person Re-identification with Part-Aware Transformer

Yulin Li[1*], Jianfeng He[1*], Tianzhu Zhang[1†], Xiang Liu[2], Yongdong Zhang[1], Feng Wu[1]
[1] University of Science and Technology of China [2] Dongguan University of Technology
{liyulin, hejf}@mail.ustc.edu.cn {tzzhang, fengwu, zhyd73}@ustc.edu.cn
succeedpkmba2011@163.com

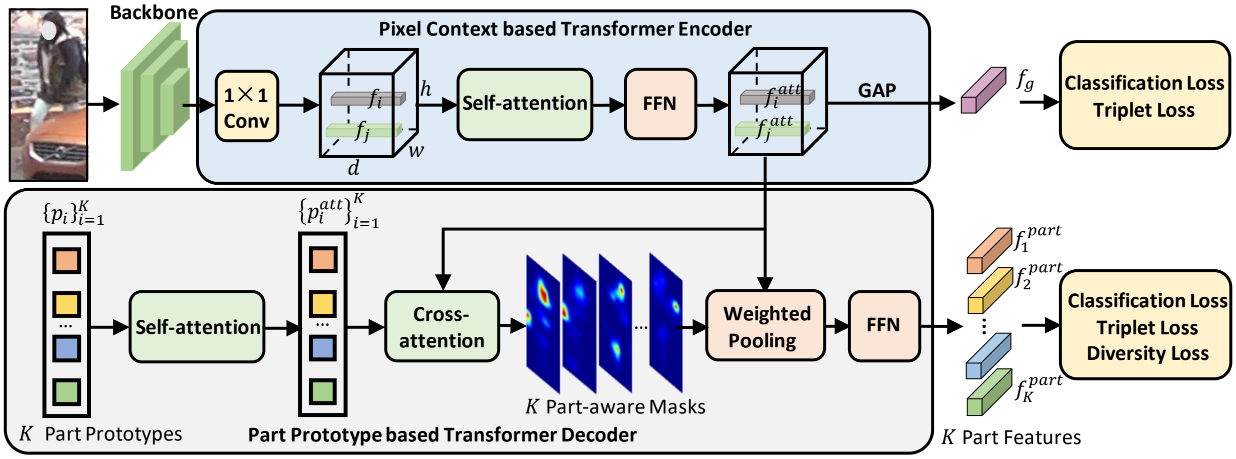

Figure 1: The pipeline of the proposed PAT consist of a pixel context based transformer encoder and a part prototype based transformer decoder. Here, "self-attention" denotes the self-attention layer, "cross-attention" denotes the cross-attention layer, and "FFN" denotes the feed forward network.

In the supplementary material, we first introduce the details about the multi-head attention mechanism and the feed-forward network. Then, we present the hyperparameters on the six datasets respectively, and then show more visualization results. Finally, we give some discussions about the difference between our method and three relevant holistic Re-ID methods.

1. Model Architecture

The overall architecture of the proposed model is shown in Figure 1, which consists of a pixel context based transformer encoder and a part prototype based transformer decoder. The detailed architecture of the self-attention layer is shown in Figure 2 (a). Since our model is based on the Transformer architecture [4, 2], in this section, we give more details about the multi-head attention mechanism and the feed-forward network.

1.1. Multi-head Attention Mechanism

In the paper, we introduce the main process of the attention mechanism. Here, we introduce the implementation of the multi-head attention mechanism. For example, in our pixel based transformer encoder, the output of the self-attention mechanism is defined as a weighted sum over all values according to the attention weights:

$$\hat{f}_i^{att} = \mathrm{Att}\left(\mathbf{Q}_i, \mathbf{K}, \mathbf{V}\right) = \sum_{j=1}^{hw} s_{i,j} \mathbf{V}_j, \tag{1}$$

We rewrite Eq.(1) for the head n as:

$$\hat{f}_{i,n}^{att} = \mathrm{Att}\left(\mathbf{Q}_{i,n}, \mathbf{K}_n, \mathbf{V}_n\right) = \sum_{j=1}^{hw} s_{i,j}^n \mathbf{V}_{j,n}, \tag{2}$$

where $n \in 1, 2, \ldots, H$ and H is the number of heads. We set H as 8 in all our experiments. The multi-head attention

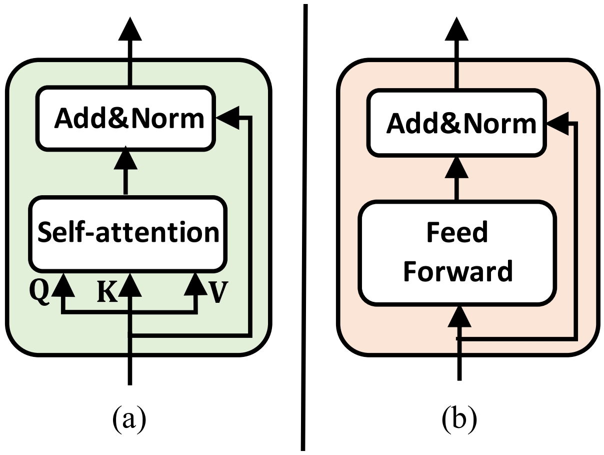

Figure 2: (a) The detailed architecture of the self-attention layer in the pipeline of the proposed PAT. (b) The detailed architecture of the feed-forward network in the pipeline of the proposed PAT.

is the concatenation of H single attention heads followed by a projection. And residual connections and layer normalization are also applied as the Transformer architecture [4, 2]. Formally,

$$\begin{aligned} \tilde{f}_i^{att} &= \mathrm{concat}(\hat{f}_{i,1}^{att}, \hat{f}_{i,2}^{att}, \ldots, \hat{f}_{i,H}^{att})\mathbf{W}^O, \\ \hat{f}_i^{att} &= \mathrm{LayerNorm}(f_i + \tilde{f}_i^{att}), \end{aligned} \tag{3}$$

where $\mathbf{W}^O \in \mathbb{R}^{Hd_v \times d}$ and LayerNorm is the layer normalization [1]. The multi-head attention mechanism in the part prototype based transformer decoder is similar to the process described above. Notice that the visualized part-aware masks in our paper are the average of masks from all attention heads of the cross-attention layer.

1.2. Feed-Forward Network

The feed forward network is a simple neural network composed of two fully connected layers with ReLU activations. Formally, given $x \in \mathbb{R}^{l_x \times d}$ as the input:

$$\mathrm{FFN} = \mathrm{ReLU}\left(x\mathbf{W}_1 + b_1\right)\mathbf{W}_2 + b_2, \tag{4}$$

where $\mathbf{W}_1 \in \mathbb{R}^{d \times d}, \mathbf{W}_2 \in \mathbb{R}^{d \times d}, b_1 \in \mathbb{R}^{1 \times d}, b_2 \in \mathbb{R}^{1 \times d}$. Besides, a residual connection followed by a layer normalization is also applied. The detailed architecture of the feed-forward network is shown in Figure 2 (b).

2. Hyperparameters and More Results

In this section, we present the hyperparameters used for each dataset and give more visualization results to analyze the effectiveness our method.

2.1. On Two Occluded Datasets

- The margin for the triplet loss α is set to 0.3.
- The number of part prototype K is set to 14.
- The loss coefficient λ_{cls} of the classification loss is set to 0.3.
- The loss coefficient λ_{tri} of the triplet loss is set to 1.
- The loss coefficient λ_{div} of the diversity loss is set to 1.

2.2. On Two Partial Datasets

- The margin for the triplet loss α is set to 0.3.
- The number of part prototype K is set to 14.
- The loss coefficient λ_{cls} of the classification loss is set to 0.3.
- The loss coefficient λ_{tri} of the triplet loss is set to 1.
- The loss coefficient λ_{div} of the diversity loss is set to 1.

2.3. On Market-1501 Dataset

- The margin for the triplet loss α is set to 0.3.
- The number of part prototype K is set to 6.
- The loss coefficient λ_{cls} of the classification loss is set to 1.
- The loss coefficient λ_{tri} of the triplet loss is set to 1.
- The loss coefficient λ_{div} of the diversity loss is set to 1.

2.4. On Duke-MTMC Dataset

- The margin for the triplet loss α is set to 0.3.
- The number of part prototype K is set to 14.
- The loss coefficient λ_{cls} of the classification loss is set to 1.
- The loss coefficient λ_{tri} of the triplet loss is set to 1.
- The loss coefficient λ_{div} of the diversity loss is set to 1.

2.5. More Visualization Results

In order to better understand the effectiveness of our method, we show more visualization results of discovered parts in Figure 3. We choose part-aware masks obtained from five part prototypes as the representative results. We also show some failure cases of the proposed method, which is shown in Figure 3 (b). In most cases, the part-aware masks learned from our model can precisely focus on discriminative human regions, which can be seen in Figure 3 (a). However, when the target person is seriously occluded by obstacles or is occluded by another person, our model tends to make mistakes and focus on all the pedestrians in the image. Since there are no precise annotations for the target person, our model cannot identify which one is our target when the target person is occluded by another person. To deal with this issue, it would be better to model the human structure information. This is what we need to solve in the future.

We also show some retrieval results of the proposed PAT in Figure 4 and some failure retrieval results in Figure 5. We can see that the proposed PAT can well overcome the occlusion problem including part occlusions and human pose variations. We can also observe the benefit of identifying the personal belongings from Figure 4. However, when the target person is occluded by another person in the gallery, the proposed PAT tends to make mistakes, which is consistent with our visualization results of part-aware masks in Figure 3 (b).

3. Discussions

In this session, we show the difference among our PAT model and three relevant holistic person Re-ID methods including CAMA [5], ABD-Net [3] and ISP [6]. (1) In CAMA [5], it adopts a multi-branch network to extract different visual features, and it constrains the class activation maps (CAM) of different branches to make them focus on different discriminative regions, while our model aims to extract diverse human part features from the backbone feature map with part-aware masks. To achieve this goal, a part prototype based transformer decoder is designed to "decode" part features in our model. Our main innovation is to obtain part features with a transformer architecture, while CAMA uses a multi-branch model to explore different visual features. (2) In ABD-Net [3], it integrates channel and spatial attention modules and diversity regularizations throughout the entire network to learn discriminative global features for the holistic Re-ID task. It aims to reduce the feature correlations to encourage the learning of informative and diverse global features. Different from them, we aim to discover diverse and discriminative part features with the proposed part prototype, and we design two effective mechanism to learn these part prototypes in a weakly supervised manner. Our main idea is to utilize part features to enhance the representations of the pedestrian in the case of occlusions. (3) In ISP [6], it does clustering on feature maps to generate pseudo labels of human parts to supervise the part estimation. Compared with ISP, our model exploits a more elegant transformer encoder-decoder architecture to obtain part features, while ISP utilizes a linear layer as the classifier for part prediction. As shown in our experiment results, our method with the ResNet-50 backbone can achieve better performance on Occluded-Duke dataset than ISP, which adopts HRNet-W32 as its backbone and consumes much more time in training because of the clustering process. These results demonstrate the effectiveness of our transformer encoder-decoder architecture and two learning mechanisms for occluded person Re-ID task.

References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[2] 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, 2020.

[3] Tianlong Chen, Shaojin Ding, Jingyi Xie, Ye Yuan, Wuyang Chen, Yang Yang, Zhou Ren, and Zhangyang Wang. Abd-net: Attentive but diverse person re-identification. In *CVPR*, pages 8351–8361, 2019.

[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017.

[5] Wenjie Yang, Houjing Huang, Zhang Zhang, Xiaotang Chen, Kaiqi Huang, and Shu Zhang. Towards rich feature discovery with class activation maps augmentation for person re-identification. In *CVPR*, pages 1389–1398, 2019.

[6] Kuan Zhu, Haiyun Guo, Zhiwei Liu, Ming Tang, and Jinqiao Wang. Identity-guided human semantic parsing for person re-identification. In *ECCV*, 2020.

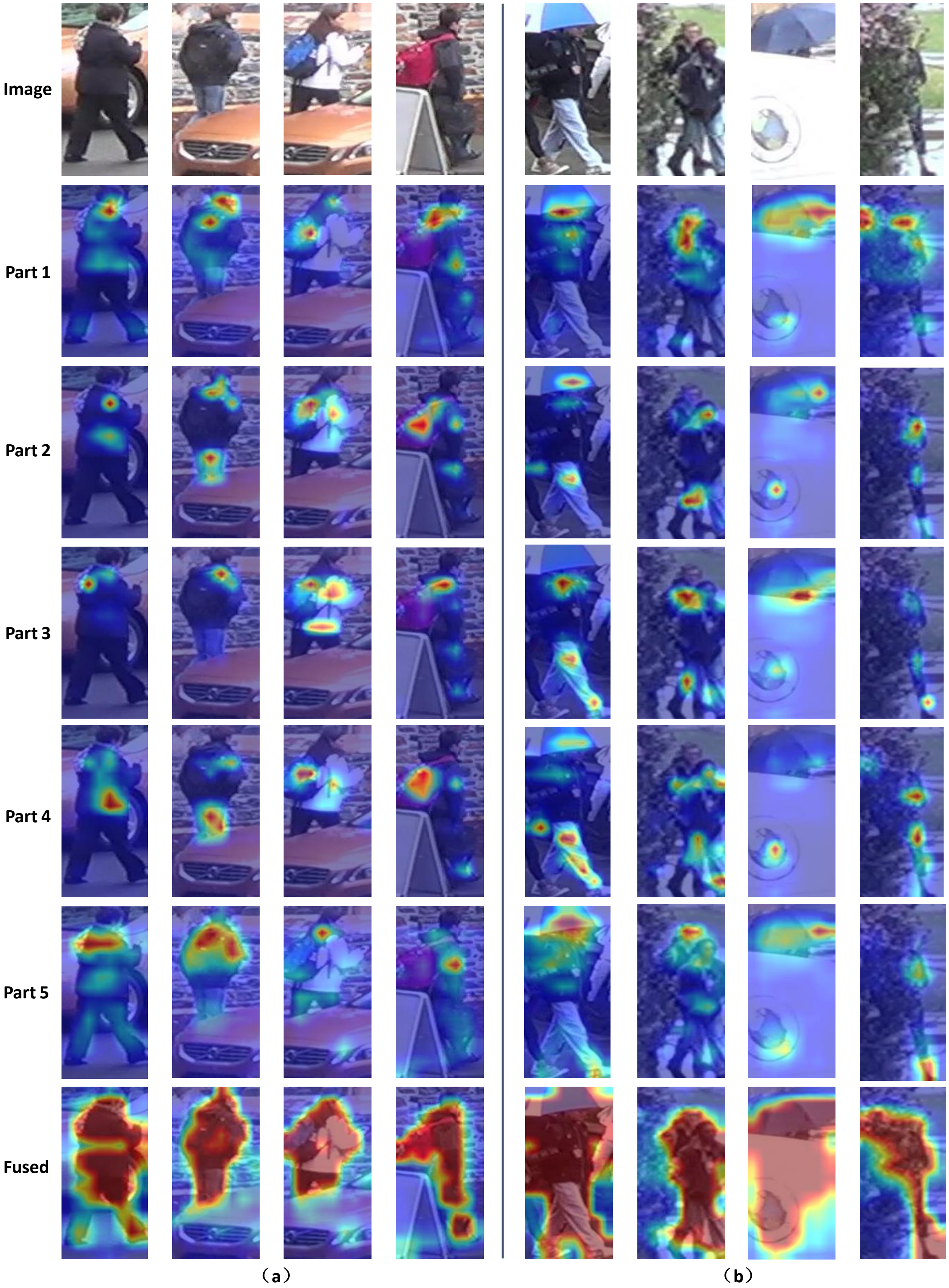

Figure 3: Visualization of the learned part-aware masks. (a) In most cases, the part-aware masks learned from our model can precisely focus on discriminative human regions. (b) When the target person is seriously occluded by obstacles or is occluded by another person, our model tends to make mistakes. This is what we need to solve in the future.

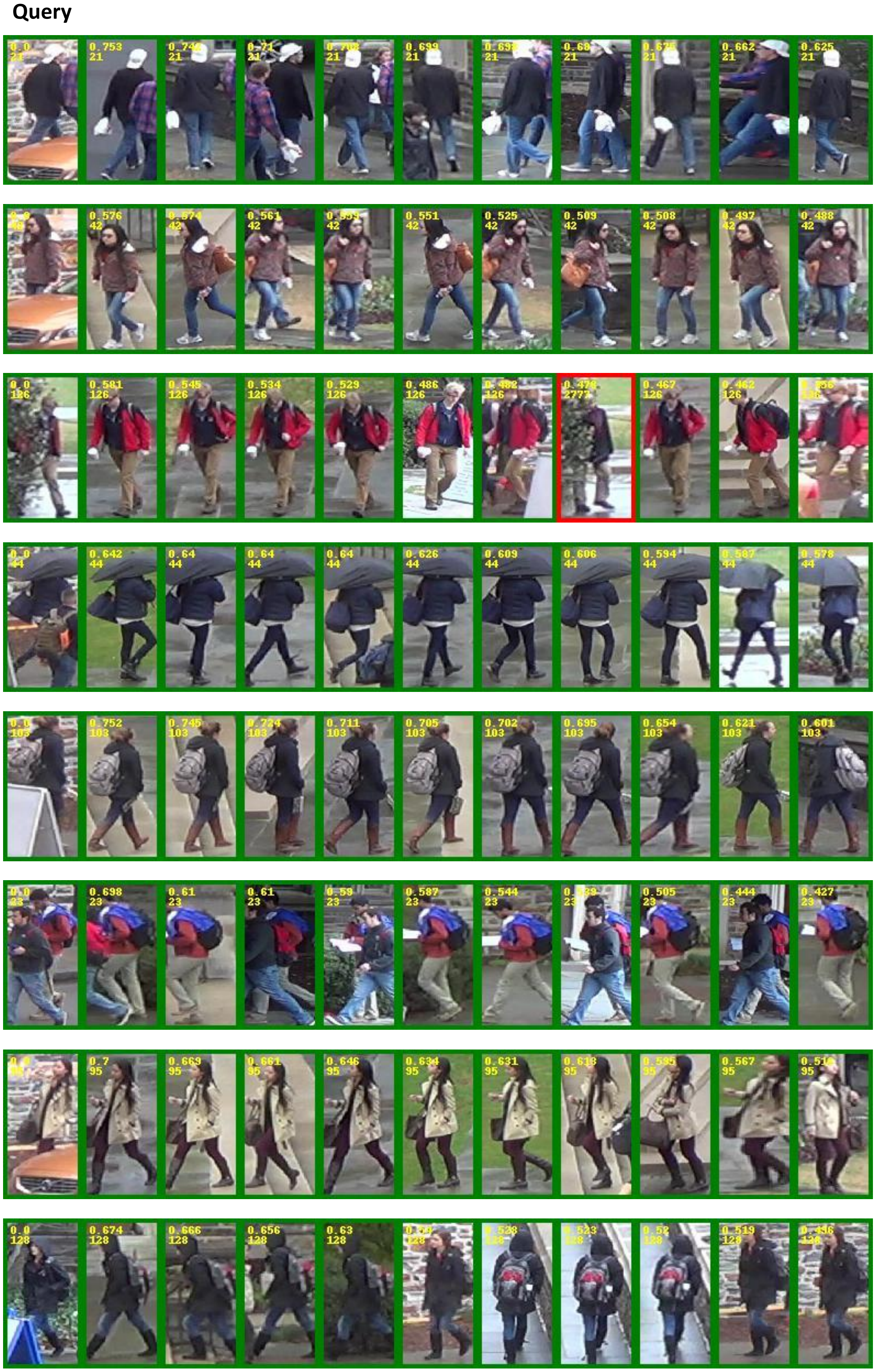

Figure 4: The retrieval results of the proposed PAT, which indicates PAT can well overcome the occlusion problem. The first image of each line is the query image. The green rectangles indicate correct retrieval results and red rectangles denote false retrieval results. The numbers in the upper left corner of the image show the cosine distance between query and gallery images and identity labels.

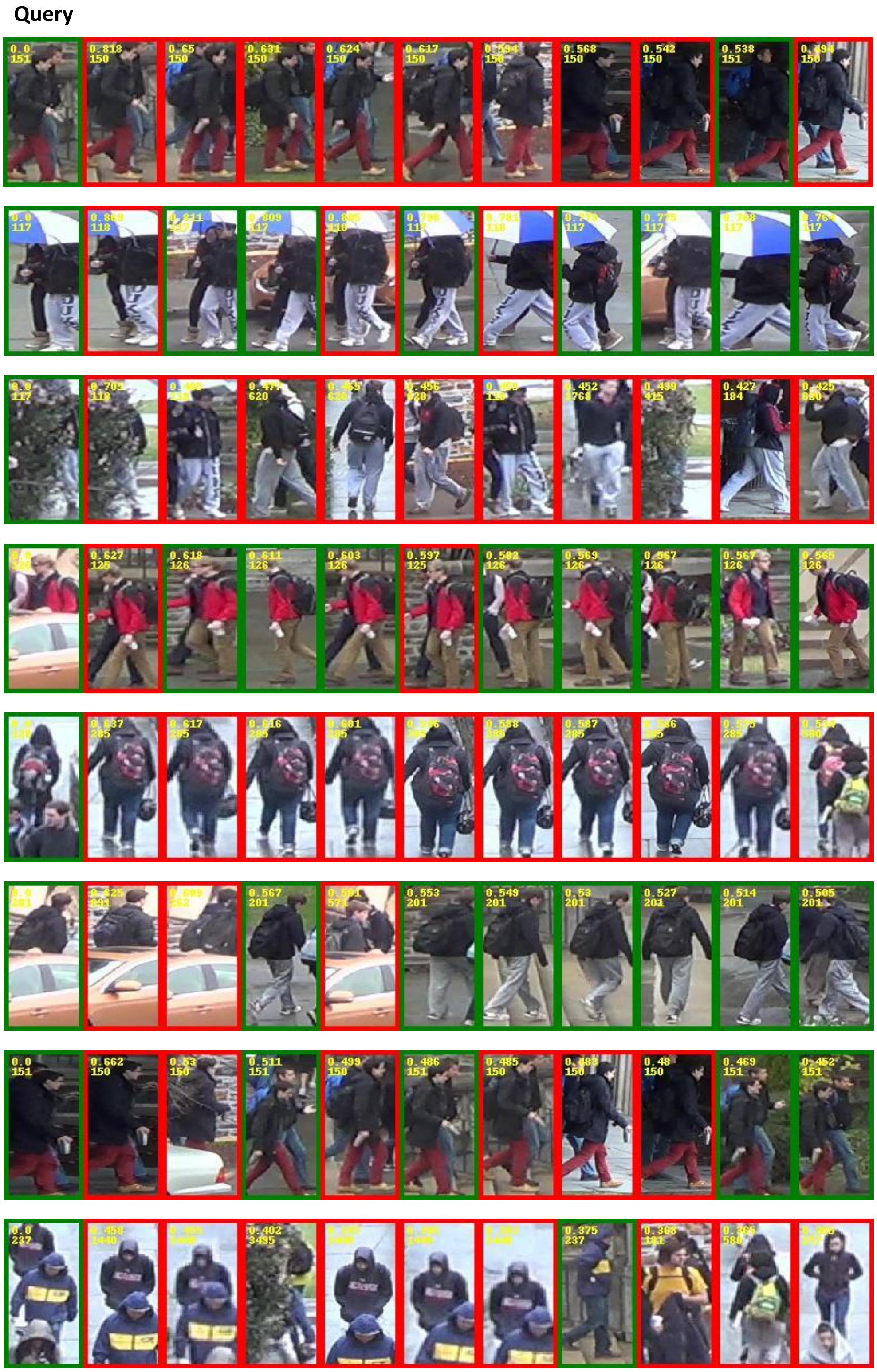

Figure 5: Some failure retrieval results of the proposed PAT. When the target person is occluded by another person in the gallery, the proposed PAT tends to make mistakes. The first image of each line is the query image. The green rectangles indicate correct retrieval results and red rectangles denote false retrieval results. The numbers in the upper left corner of the image show the cosine distance between query and gallery images and identity labels.