Part Representation Learning with Teacher-Student Decoder for Occluded Person Re-Identification

Shang Gao[1,2], Chenyang Yu[1], Pingping Zhang[1*], and Huchuan Lu[1,2]

1 Dalian University of Technology, Dalian, China

2 NingBo Institute of Dalian University of Technology, Ningbo, China

gs940601k@gmail.com, zhpp@dlut.edu.cn, lhchuan@dlut.edu.cn

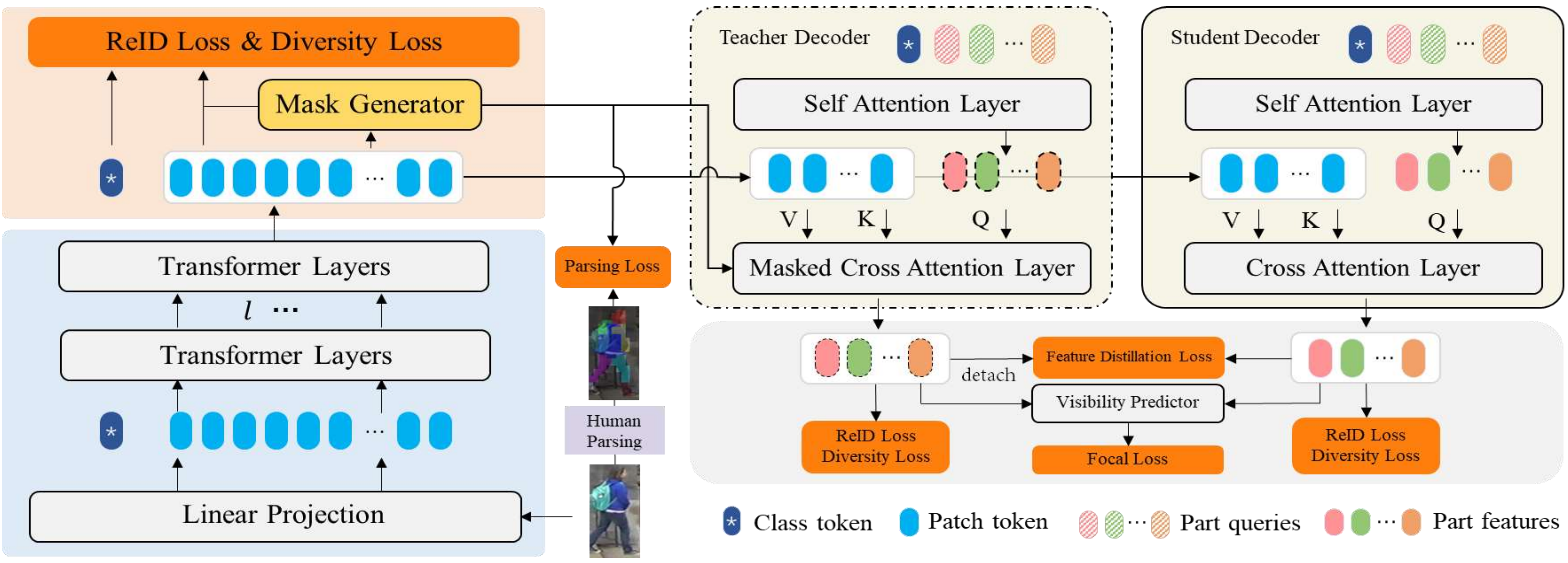
Motivation and Contribution

- Disentangling body part features and conducting part-to-part comparison on the visible body parts is a mainstream solution for occluded person ReID.
- Transformer encoder-decoder model has shown powerful capabilities in many vision tasks. However, it fails in adequately disentangling body part features with merely global supervision for person ReID
- Leveraging external cues such as human pose or parsing to locate and align part features has been proven to be very effective in occluded person ReID.
- We propose a Teacher-Student Decoder (TSD) framework to incorporate the human parsing information into the Transformer for occluded person ReID.

Re-Occluded-Duke Benchmark

- Existing occluded person ReID benchmarks utilize occluded samples as queries, which will amplify the role of alleviating occlusion interference and underestimate the impact of the feature absence issue.
- We propose a new benchmark with non-occluded queries, wherein positive holistic samples are ignored in the ranking list.

Part Representation Learning with Teacher-Student Decoder



- Cross-attention machine in Standard Student Decoder (SSD)

$$X^s = \mathrm{Softmax}(QK/\sqrt{D})V$$

- Cross-attention machine in Parsing-aware Teacher Decoder (PTD)

$$X_p^t = \mathrm{Softmax}(H_p + Q_p K)V$$

- **Feature distillation loss** to transfer knowledge from PTD to SSD:

$$L_{fd} = \frac{1}{P}\sum_i (1 - Sim(F_i^{sd}, F_i^{td}))$$

- **Mask Generation** to preserve model from noisy parsing results

$$M = Softmax(F^{pt}G^T)$$

$$L_m = L_{ce}(F_c^{part}) + L_{tri}^p(F^{part}) + L_{pa}(M)$$

- **Diversity loss** to preserve model from extracting identical features

$$L_{div} = \frac{1}{P(P-1)} \sum_{i=1, i\neq j}^{P} \sum_{j=1}^{P} Sim(F_i^{td}, F_j^{td})$$

Experiment and Visualization

Table 2. Comparison with other methods on our benchmark.

Method	OCC		NPO		NTP	
	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
ViT-B [1]	67.1	52.5	60.8	51.1	60.1	51.4
FED [23]	63.9	47.4	57.6	46.0	56.7	46.6
BPBreID* [6]	67.8	54.1	61.5	53.4	59.0	50.4
DPM [24]	69.2	53.5	62.0	50.8	63.6	53.9
PFD [9]	70.9	55.7	64.8	54.3	**64.6**	**55.2**
Ours	**71.4**	**58.7**	**68.0**	**61.5**	61.9	52.5
SAP* [14]	71.4	57.1	65.8	55.4	**65.4**	56.6
Ours *	**73.2**	**61.7**	**68.8**	**62.7**	64.9	**57.5**

Table 1. Comparison with other state-of-the-art methods on Occluded-Duke and DukeMTMC-reID. ∗ indicates the backbone is with an overlapping stride setting. † indicates it is reproduced by replacing the original backbone with ViT.

Method	Occluded-Duke		DukeMTMC-reID	
	Rank-1	mAP	Rank-1	mAP
ViT-B [1]	61.5	53.5	88.8	79.3
TransReID [22]	64.2	55.7	89.6	80.6
BPBreID† [6]	66.0	56.7	90.2	80.8
PFD [9]	67.7	**60.1**	**90.6**	**82.2**
FED [23]	68.1	56.4	89.4	78.0
Ours	**70.6**	57.3	90.2	81.7
DPM* [24]	71.4	61.8	91.0	82.6
SAP* [14]	70.0	62.2	-	-
PFD* [9]	69.5	61.8	**91.2**	**83.2**
Ours *	**74.5**	**62.8**	90.8	82.8

Table 3. Ablation study for the main components on Occluded-Duke and our benchmark.

Method	Occluded-Duke		OCC	NPO	NTP
	Rank-1	mAP	mAP	mAP	mAP
Baseline	61.5	53.5	52.3	50.2	50.8
M1	59.9	51.5	49.0	47.4	48.4
M2	65.1	54.3	54.6	57.8	48.6
M3	68.3	54.9	54.5	56.8	48.2
M4	70.6	57.3	58.7	61.5	52.5

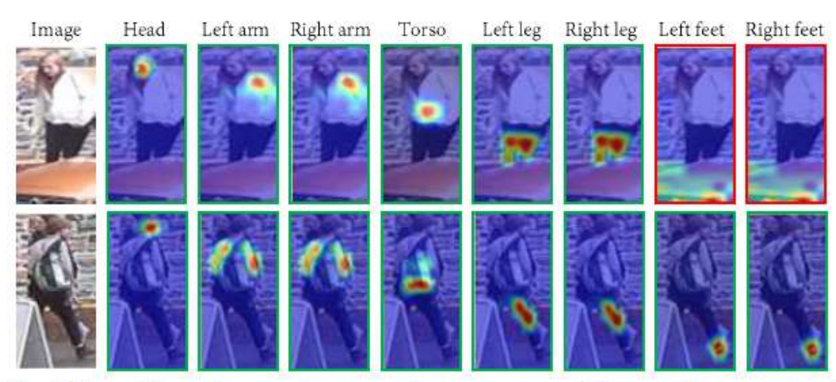


Fig. 3. Visualization of attention maps. Green boxes indicate visible predictions, while red boxes for invisible ones.