View-decoupled Transformer for Person Re-identification under Aerial-ground Camera Network

Supplementary Material

7. Experiments

7.1. Setting

VDT adopts the ViT-base [34] as the baseline, which contains $N = 12$ encoder blocks and is pre-trained on the ImageNet [45]. The patch size and stride size in VDT are set to 16×16. The size of the input image is resized to 256×128, so the $M = 128$. The embedding shape d of tokens is set to 768. The t_m and t_v are randomly initialized at the beginning of training. During training, we adopt padding with 10 pixels, random cropping, and random erasing with a probability of 0.5 for data augmentation. We adopt a soft version of triplet [21] to avoid manually selecting m in the triplet loss. The stochastic gradient descent [46] optimizer is used. The cosine learning rate decay is adopted to reduce the learning rate from initial 8×10^{-3} to final 1.6×10^{-6}. The number of training epochs is 120. The batch size is 128, including 32 identities, each with four images. We do not apply any data augmentation or re-ranking during inference. The VDT is implemented by PyTorch [36, 47]. All experiments have been conducted on one A5000 GPU.

7.2. Visualization

Retrieval visualization. Fig. 6 shows the retrieval advantages of VDT under multiple protocols of the two datasets. Compared to the baseline, VDT achieves better feature decoupling and extracts more discriminative descriptions of the target person from view-unrelated features, which makes the identity features more robust under each protocol of both datasets. Fig. 6 amply illustrates that the proposed view decoupling is feasible and effective for alleviating view discrepancy in AGPReID.

Feature visualization. As shown in Fig. 7, we randomly select a pedestrian identity on each dataset and visualize the meta (circle) and view (triangles) tokens corresponding to all the images under this identity (the same color means from the same image). In Fig. 7, meta and view tokens show good cohesion and significant differences from each other, which indicates that our method achieves good decoupled representations between these two tokens.

7.3. Cross-dataset evaluation

The results (training on CARGO, and testing on AR-ReID) have been shown in Tab. 4, which indicates that the direct cross-dataset (domain) evaluation is challenging, but our VDT has advantages over the baseline. One possible reason is that our decoupling strategy allows identity-related learning to be less perturbed by domain-related factors (e.g., view bias), leading to more discriminative identity features.

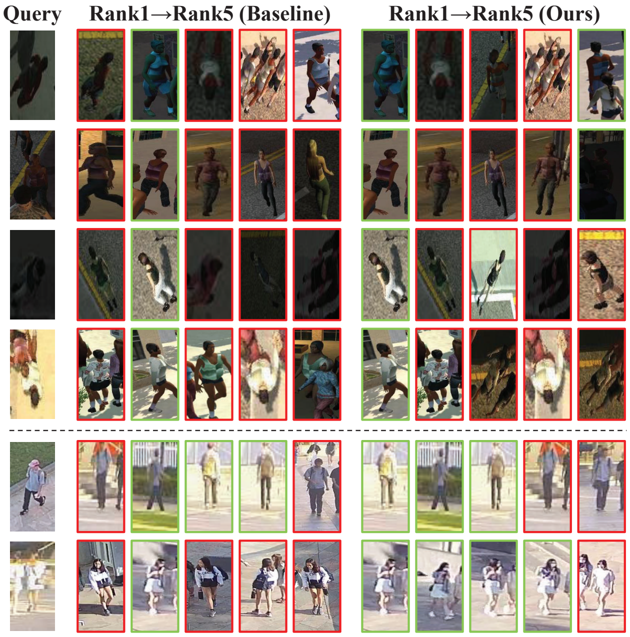
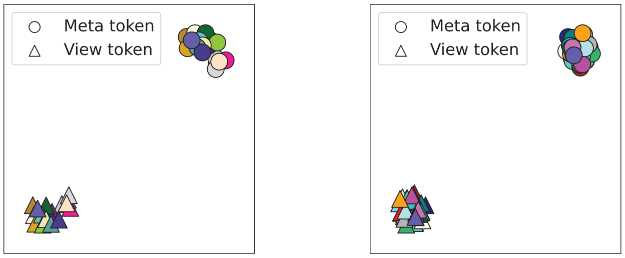


Figure 6. Comparison of several retrieval visualizations on the CARGO and AG-ReID dataset protocols. Red and green boxes represent wrong and correct matchings. The top five are listed.



(a) CARGO dataset

(b) AG-ReID dataset

Figure 7. Visualization of meta and view tokens via tSNE.

Table 4. Cross-domain performance evaluations (%) for transferring from CARGO to AG-ReID dataset.

Method	CARGO→AG-ReID					
	Protocol1: A→G			Protocol2: G→A		
	Rank1	mAP	mINP	Rank1	mAP	mINP
ViT [34]	1.59	1.95	0.8	3.01	2.31	0.95
VDT (Ours)	**19.33**	**11.81**	**1.63**	**15.38**	**11.73**	**3.38**

Table 5. Performance comparison (%) on CARGO and AG-ReID datasets.

Method	CARGO			AG-ReID			AG-ReID		
	Protocol1: ALL			Protocol1: A→G			Protocol2: G→A		
	Rank1	mAP	mINP	Rank1	mAP	mINP	Rank1	mAP	mINP
TransReID [29]	60.90	53.17	39.57	78.25	70.03	44.74	79.74	70.79	45.12
VDT (Ours)	**64.10**	**55.20**	**41.13**	**82.91**	**74.44**	**51.06**	**86.59**	**78.57**	**52.87**

7.4. Discussion

The results of TransReID [29] on two datasets have been shown in Tab. 5, which shows that show that although TransReID [29] introduces additional camera information and encodes it as part of the input, its performance on both datasets is still weaker than that of the our method, proving the effectiveness of the proposed decoupling strategy proposed for the AGPReID task.