View-Aware Semantic Alignment for Aerial-Ground Person Re-Identification

Supplementary Material

In this supplementary material, we provide additional analyses and results to further evaluate our proposed ViSA. Specifically, we first present extended qualitative visualizations, including additional retrieval examples on CARGO and t-SNE comparisons. We then report comprehensive quantitative results by providing full performance tables on AG-ReID.v2 and LAGPeR under all evaluation settings. Finally, we conduct detailed hyperparameter analyses, examining the effects of the number of experts, the number of selected experts, and the balancing coefficient across different cross-view protocols.

6. Visual Analysis

6.1. Retrieval Visualization

In addition to the retrieval visualization on the AG-ReID.v2 dataset shown in Fig. 5, we further provide retrieval examples on the CARGO dataset under both the A↔G protocol and the ALL protocol. As illustrated in Fig. 6, our method consistently retrieves correct cross-view matches across different evaluation settings, demonstrating strong robustness and generalization. These results further verify the effectiveness of our approach in handling diverse viewpoint transitions across datasets and protocols.

6.2. Feature Visualization

To further investigate the effectiveness of ViSA in cross-view representation learning, we visualize the learned embeddings using *t*-SNE. As shown in Fig. 7, we compare the feature distributions of the baseline model and ViSA. The reduced discrepancy between ground and aerial features indicates that ViSA effectively mitigates cross-view domain gaps and learns more discriminative, view-invariant representations.

7. Performance

Tab. 3 and Tab. 4 report the performance of ViSA on AG-ReID.v2 and LAGPeR, respectively. On AG-ReID.v2, ViSA achieves the highest mAP across all methods and sets a new state of the art in every evaluation setting except the A→W protocol, where the viewpoint variation is relatively minor. We further observe that in this less challenging A→W scenario, methods such as Explain and V2E obtain slightly better results, largely benefiting from the use of attribute annotations. When compared only with approaches that do not rely on attribute labels, ViSA still delivers the best performance. On the LAGPeR dataset, ViSA achieves improvements of up to 1.0% in mAP and 1.22% in Rank-1 accuracy, demonstrating its strong generalization ability under larger cross-view variations.

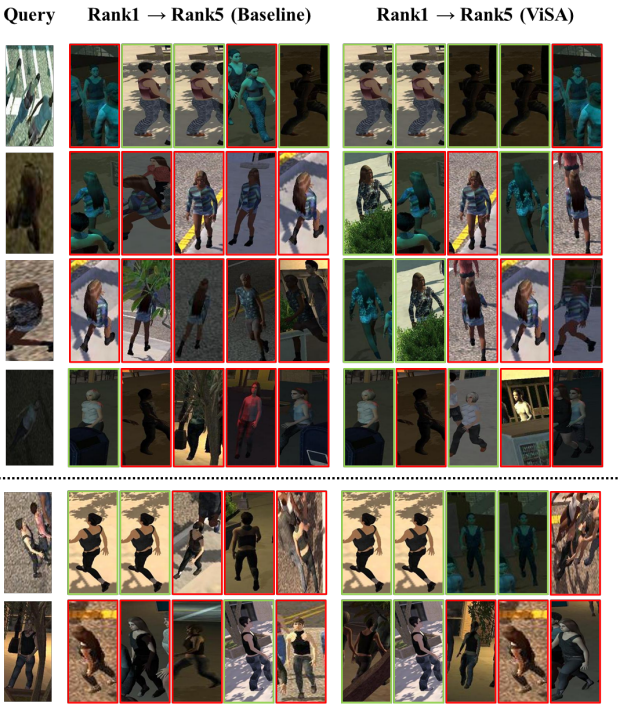


Figure 6. Comparison of several retrieval visualizations on CARGO dataset under A→G protocol and ALL protocol. Red and green boxes represent wrong and correct matchings, respectively. The top five retrieved results are shown.

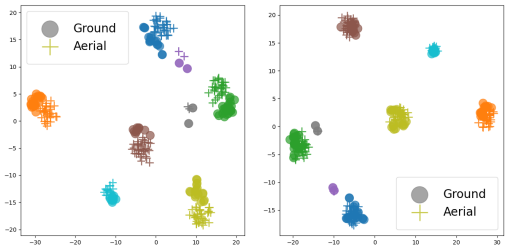


Figure 7. t-SNE visualization. **Left:** Baseline. **Right:** ViSA. Colors indicate different classes and shapes denote different views.

8. Parameter Analysis

We further analyze the impact of hyperparameters under different evaluation protocols. In the study of the number of experts E, we fix the number of selected expert to 1 (setting $k = 1$) to ensure controlled comparisons. As shown in

Table 3. Performance comparison on AG-ReID.v2 dataset. C represents CCTV, W represents wearable devices and A represents aerial views. The best and second-best results are highlighted in **bold** and underline, respectively.

Method	Venue	A→C		A→W		C→A		W→A	
		Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
Swin [49]	ICCV'21	68.76	57.66	68.49	56.15	68.80	57.70	64.40	53.90
HRNet-18 [52]	PAMI'20	75.21	65.07	76.26	66.17	76.25	66.16	76.25	66.17
SwinV2 [53]	CVPR'22	76.44	66.09	80.08	69.09	77.11	62.14	74.53	65.61
MGN(R50) [40]	ACM MM'18	82.09	70.17	88.14	78.66	84.21	72.41	84.06	73.73
BoT(R50) [41]	CVPR'19	80.73	71.49	86.06	75.98	79.46	69.67	82.69	72.41
SBS(R50) [45]	ACM MM'23	81.96	72.04	88.14	78.94	84.10	73.89	84.66	75.01
BoT(ViT) [41]	CVPR'19	85.40	77.03	89.77	80.48	84.65	75.90	84.65	75.90
ViT [46]	ICLR'21	85.40	77.03	89.77	80.48	84.65	75.90	84.27	76.59
TransReID [54]	ICCV'21	88.00	81.40	90.40	84.50	87.60	80.10	87.70	81.10
FusionReID [55]	T-ITS'25	86.70	80.70	89.70	84.20	87.90	80.00	86.50	80.90
CLIP-ReID [48]	AAAI'23	85.36	79.79	89.14	84.23	85.64	79.08	86.50	79.55
PCL-CLIP [47]	CoRR'23	79.80	72.20	87.14	77.70	81.12	72.40	84.19	73.89
Explain [15]	ICME'23	87.70	79.00	**93.67**	83.14	87.35	78.24	87.73	79.08
VDT [19]	CVPR'24	86.46	79.13	90.00	82.21	86.14	78.12	85.26	78.52
V2E(ViT) [17]	TIFS'24	88.77	80.72	93.62	84.85	87.86	78.51	88.61	80.11
SeCap [21]	CVPR'25	88.12	80.84	91.44	84.01	88.24	79.99	87.56	80.15
ViSA	-	**89.43**	**83.61**	91.63	**85.99**	**88.57**	**82.32**	**89.23**	**83.10**

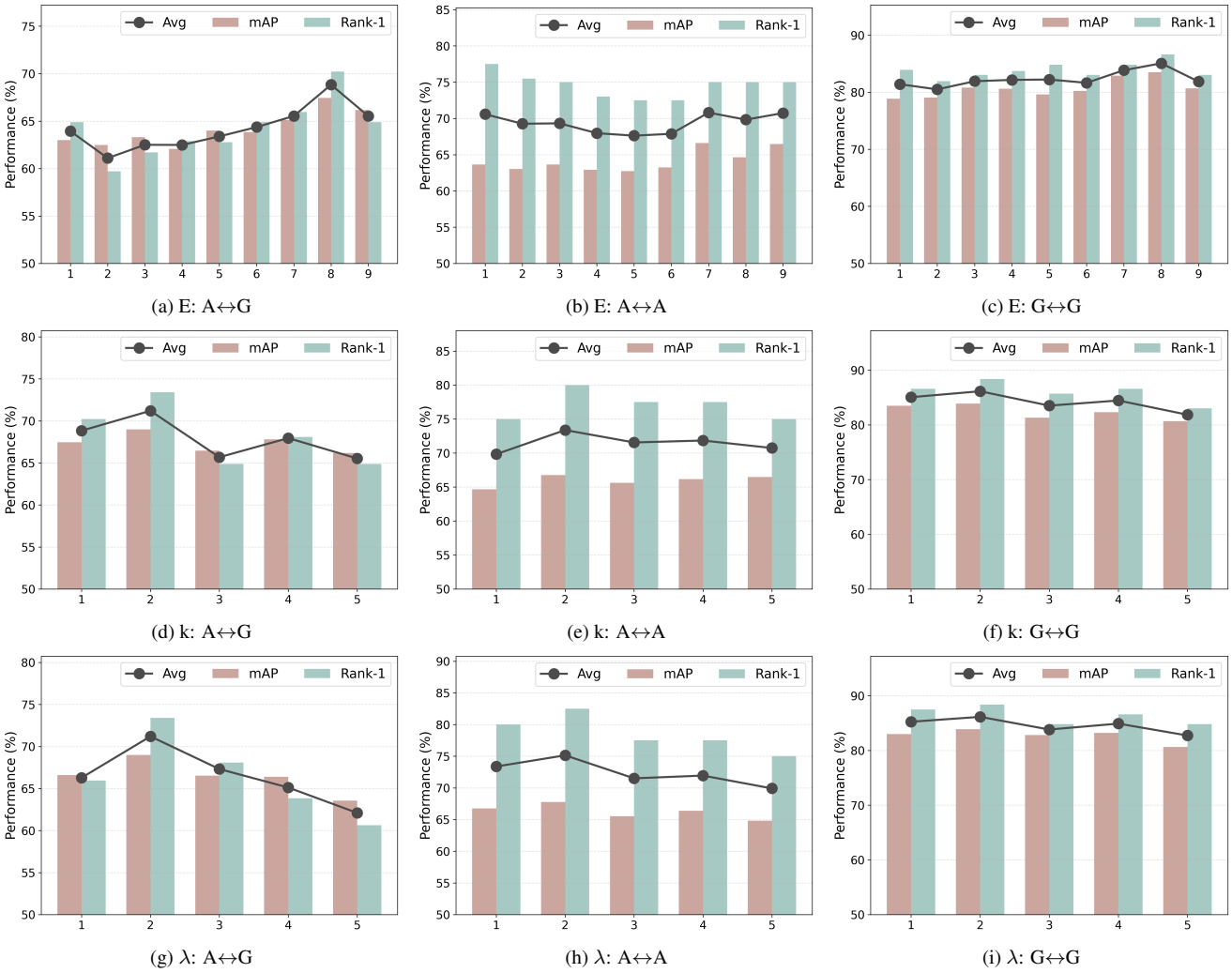
Table 4. Performance comparison on LAGPeR. The best and second-best results are highlighted in **bold** and underline, respectively. 'A↔G' denotes that the Aerial view is the query, 'G↔A' denotes that the Ground view is the query, and 'G↔A+G' indicates that the gallery contains images from both the Aerial and Ground view. CLIP-ReID* indicates using OLP and SIE in Clip-ReID. MIP† represents the re-implementation for the AGPReID. Explain‡ indicates removing the attributes branch of the Explain method.

Method	Venue	A↔G		G↔A		G↔A+G		Average	
		Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
ViT [46]	ICLR'21	38.67	27.25	32.04	30.69	18.88	15.31	29.86	24.42
TransReID [54]	ICCV'21	38.80	28.80	33.00	32.10	22.90	18.80	31.57	26.57
CLIP-ReID [48]	AAAI'23	24.40	17.60	21.30	20.80	12.30	10.20	19.33	16.20
CLIP-ReID*	AAAI'23	23.10	17.50	20.00	20.30	9.00	8.40	17.37	15.40
MIP† [56]	ICMR'24	39.30	29.30	33.90	32.60	21.00	17.30	31.40	26.40
Explain‡ [15]	ICME'23	40.48	28.89	32.96	31.91	22.03	17.89	31.82	26.23
VDT [19]	CVPR'24	40.15	28.97	33.55	31.98	19.50	16.45	31.07	25.80
SeCap [21]	CVPR'25	**41.79**	30.37	**35.26**	33.42	24.39	19.24	33.81	27.68
ViSA	-	41.37	**30.42**	35.06	**33.59**	**25.61**	**20.31**	**34.01**	**28.11**

Fig. 8a ∼ Fig. 8c, the performance first increases and then decreases as E grows. Under the A↔G and G↔G protocols, the best results are achieved when $E = 8$, while under the A↔A protocol, the optimal performance is obtained with $E = 7$.

Similarly, as shown in Fig. 8d ∼ Fig. 8f when analyzing the number of selected experts k, we observe a clear rise-then-drop trend. The performance peaks at $k = 2$, indicating that activating a small but diverse subset of experts provides the best balance between specialization and generalization.

For the balancing coefficient λ, we compare values $\{0.0001, 0.001, 0.01, 0.1, 1.0\}$ across all protocols. As shown in Fig. 8g ∼ Fig. 8i, setting $\lambda = 0.001$ consistently achieves the best performance on all three protocols, indicating that a moderate regularization strength is essential



(a) E: A↔G (b) E: A↔A (c) E: G↔G

(d) k: A↔G (e) k: A↔A (f) k: G↔G

(g) λ: A↔G (h) λ: A↔A (i) λ: G↔G

Figure 8. Analysis of E, k, and λ across different protocols on the CARGO dataset. Rank-1 and mAP and their average are reported (%). E represents the number of experts. k represents number of selected experts. λ represents the balancing coefficient.

Table 5. Efficiency comparison under ALL protocol.

Method	mAP(↑)	Params(↓)	GFLOPs(↓)	FPS(↑)
SeCap	58.94%	130.88M	38.81	276.66
ViSA	69.00%	271.77M	37.53	429.05
Rel.	**+10.06%**	+107.6%	**-3.3%**	**+55.1%**

for maintaining stable expert utilization without suppressing discriminative specialization.

9. Efficiency

ViSA achieves strong accuracy and efficiency despite introducing additional parameters (+107.6%). As shown in Tab. 5, ViSA enhances mAP by 10.06% while reducing GFLOPs by 3.3% and increasing inference speed by 55.1% FPS.