SeCap: Self-Calibrating and Adaptive Prompts for Cross-view Person Re-Identification in Aerial-Ground Networks

Supplementary Material

7. Overview

In this supplementary material, we provide additional experimental results and more in-depth discussions of the following three aspects:

- We conduct a visual analysis comparing our proposed SeCap method with the baseline model, including retrieval results and attention map.
- We perform experiments on the AGPReID datasets GARGO [38] and AG-ReID.v2 [21], demonstrating that SeCap is a feasible and effective solution for the AGPReID task across all publicly available AGPReID datasets. Additionally, we conducted cross-dataset evaluation experiments.
- We analyze the impact of prompt length L and hyperparameter λ on model performance and conduct ablation experiments on the individual modules to verify the effectiveness of our method.

Unless otherwise specified, the numbering of figures and tables should be within the scope of the supplementary material, and consistent with the main paper.

8. Visual Analysis

8.1. Retrieval Result Visualization

The visualization of the retrieval results compellingly demonstrates that SeCap is a feasible and effective method for addressing the challenges posed by AGPReID problems. As illustrated in Fig. 4, the retrieval outcomes on both LAGPeR and AG-ReID datasets are presented, offering a comprehensive comparison of SeCap's retrieval results with those of baseline methods across various experimental settings.

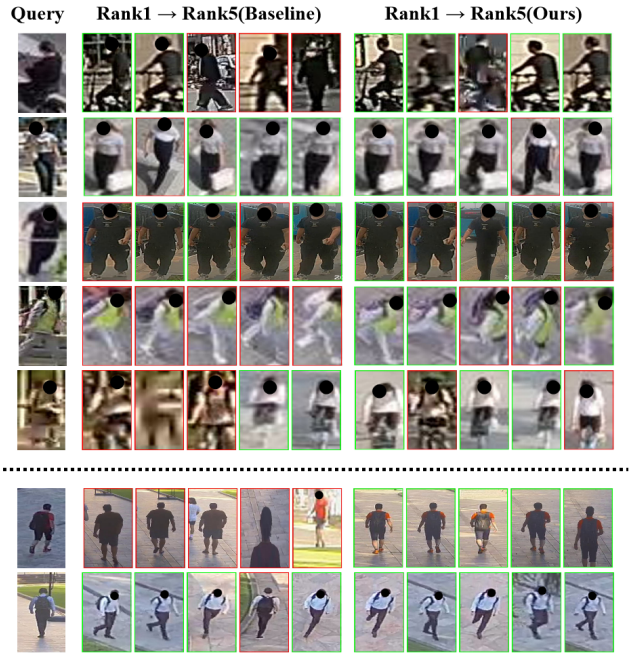


Figure 4. Comparison of several retrieval visualizations on the LAGPeR dataset of setting $A \rightarrow G$. Red and green boxes represent wrong and correct matchings. The top five are listed.

8.2. Attention Map Visualization

We visualized the attention maps of some case images from both the SeCap method and the baseline model. As shown in Fig. 5, the baseline model tends to focus more on the torso or clothing of individuals rather than view-invariant regions like head features. **In contrast, our SeCap method effectively attends to view-invariant local features, ensuring robust performance across varying viewpoints**.

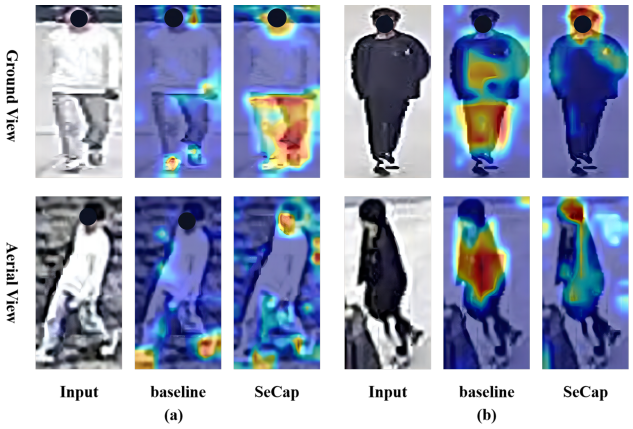


Figure 5. The visualization results of the attention maps of our SeCap method and the baseline model.

9. Performance on Other AGPReID Datasets

9.1. Dataset.

(1) AG-ReID.v2: This dataset comprises $100,502$ images with $1,605$ unique IDs, captured by three types of cameras: CCTV, UAV, and wearable devices [21]. Among these, 807 IDs are designated for the training set, while the remaining 798 IDs are used for the test set. Additionally, the dataset includes 15 attributes to facilitate cross-view matching. In terms of experimental settings, the test images are evaluated

Table 6. Performance comparison under CARGO dataset. 'ALL' denotes the overall retrieval performance of each method. '$G \leftrightarrow G$', '$A \leftrightarrow A$', and '$A \leftrightarrow G$' represent the performance of each model in several specific retrieval patterns.Rank1 and mAP are reported (%). The best performance is shown in **bold**.

METHOD	BACKBONE	ALL		$G \leftrightarrow G$		$A \leftrightarrow A$		$A \leftrightarrow G$	
		Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
SBS [6]	R50	50.32	43.09	72.31	62.99	67.50	49.73	31.25	29.00
AGW [34]	R50	60.26	53.44	81.25	71.66	67.50	56.48	43.57	40.90
BoT [17]	ViT	61.54	53.54	82.14	71.34	80.00	64.47	43.13	40.11
VDT [38]	ViT	64.10	55.2	82.14	71.59	**82.50**	66.83	48.12	42.76
SeCap(Ours)	ViT	**68.59**	**60.19**	**86.61**	**75.42**	80.00	**68.08**	**69.43**	**58.94**

Table 7. Performance comparison on the AG-ReID.v2 dataset. C represents CCTV, W represents wearable devices, and A represents aerial views. The best results are highlighted in **bold**, while the second-best results are underlined.

METHOD	BACKBONE	$A \rightarrow C$		$C \rightarrow A$		$A \rightarrow W$		$W \rightarrow A$	
		Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
BoT [17]	ViT	85.40	77.03	84.65	75.90	89.77	80.48	84.65	75.90
AG-ReIDv1 [20]	ViT	87.70	79.00	87.35	78.24	**93.67**	83.14	87.73	79.08
VDT [38]	ViT	86.46	79.13	86.14	78.12	90.00	82.21	85.26	78.52
AG-ReIDv2 [21]	ViT	**88.77**	80.72	87.86	78.51	93.62	**84.85**	**88.61**	80.11
SeCap(Ours)	ViT	88.12	**80.84**	**88.24**	**79.99**	91.44	84.01	87.56	**80.15**

under the following conditions: $A \rightarrow C$, $C \rightarrow A$, $A \rightarrow W$, and $W \rightarrow A$.

(2) CARGO: The CARGO dataset is a virtual AGPReID dataset constructed using tools such as MakeHuman [1] and Unity3D. It comprises $108,563$ images with $5,000$ unique IDs, captured by 13 cameras: 8 ground cameras and 5 aerial cameras. Among these, $51,451$ images from $2,500$ IDs are designated for the training set, while the remaining 51,024 images from $2,500$ IDs are used for the test set. In terms of experimental settings, the test images are evaluated under four conditions: ALL, $A \leftrightarrow A$, $G \leftrightarrow G$, and $A \leftrightarrow G$. The "ALL" setting focuses on comprehensive retrieval performance, while the latter targets specific retrieval scenarios.

9.2. Performance.

On additional AGPReID datasets, SeCap demonstrates robust performance. Tab. 7 and Tab. 6 present the performance of the proposed SeCap on the AG-ReID.v2 [21] and CARGO [38] datasets. It can be seen that **SeCap achieves optimal results across various settings on the synthetic AGPReID dataset CARGO and significantly outperforms other methods in the cross-view task $A \leftrightarrow G$, demonstrating the significant advantages of our proposed method in solving cross-view problems**. In the setting $A \leftrightarrow A$, due to the limited number of queries in CARGO, which consists of only 60 IDs with 134 images, the chance level is 2.5%. Consequently, the Rank1 performance is relatively close. However, when considering the metric of mAP metric, which better reflects the model's performance, our method demonstrates superior results.

On the AG-ReID.v2 dataset, we compare AGPReID methods such as AG-ReID.v1, VDT, and AG-ReID.v2. AG-ReID.v1 only reports results using ResNet-50 as the backbone on the AG-ReID.v2 dataset, so we compare the results of ViT enhanced by the Explainable ReID Stream(EP). As shown in Tab. 7, we observe that **even without using the attributes provided by AG-ReID.v2, our method still achieved the best or comparable results in the $A \rightarrow C$ and $C \rightarrow A$ experimental settings. In the $A \rightarrow W$ and $W \rightarrow A$ settings, our method achieves the best or comparable mAP results, but its Rank-1 metric is not as high as AG-ReID.v2.** This discrepancy arises because our SeCap method uses view-invariant local features for matching, with head information being a significant view-invariant feature. From Fig. 5, it is evident that our method implicitly trains the model to

Table 8. The analysis of the effectiveness of the PRM and LFRM in SeCap. LFRM stands for the Local Feature Refinement Module, PRM denotes the Prompt Re-calibration Module, and OLP represents Overlapping Patches. The meanings of Add, Cat, and Attn are detailed in the Sec. 11. The best performance and the most significant improvements are highlighted in **bold**.

No.	PRM			LFRM			OLP	$A \rightarrow G$		$G \rightarrow A$		$G \rightarrow A+G$	
	Add	Cat	Attn.	Block	Two-Way	fusion		Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
1	✔				✔	✔	✔	38.48	27.76	32.14	30.49	22.75	18.49
2		✔			✔	✔	✔	40.09	29.10	33.88	32.71	22.52	18.44
3			✔	✔		✔	✔	39.92	28.59	33.32	31.55	22.88	18.60
4			✔		✔		✔	40.25	28.88	34.11	32.25	21.14	17.23
5			✔		✔	✔		40.87	29.11	33.72	32.48	19.67	16.40
6			✔		✔	✔	✔	**41.8**	**30.4**	**35.3**	**33.4**	**24.39**	**19.24**

focus more on head features. Conversely, AG-ReID.v2's Elevated-View Attention Stream(EVA) explicitly uses head information for cross-view matching, which is generally more robust than implicitly extracting local features, resulting in better Rank-1 performance. However, this approach may fail when the head is occluded, leading to a significant performance drop. Therefore, our method performs better on the average mAP metric, which better indicates the model's re-identification capability [42]. Additionally, in the $A \rightarrow W$ setting, we found that the improvement in model performance is mainly due to the attribute-based Explainable ReID Stream(EP), rather than the Elevated-View Attention Stream (EVA), which has the limitation of relying on attribute labels.

10. Cross-dataset evaluation

The proposed SeCap method in this study demonstrates superiority over other methods in cross-dataset evaluation. Specifically, as shown in Tab. 9, the results of training on the LAGPeR dataset and testing on the AR-ReID dataset indicate that direct cross-dataset (or cross-domain) evaluation is a challenging task. However, the SeCap method exhibits more significant advantages compared to baseline methods and the VDT method. This advantage may stem from the dynamically generated and calibrated prompt mechanism of SeCap, which not only learns perspective-irrelevant features but also effectively guides the model to focus more on cross-domain identity discrimination features, thereby promoting the model to learn more discriminative feature representations.

11. Effectiveness Analysis of the Modules

As shown in Tab. 8, we analyze the roles of the Prompt Re-calibration Module (**PRM**), Local Feature Refinement Module (**LFRM**), and Overlapping Patches(**OLP**).

Table 9. Cross-dataset performance evaluations (%) for transferring from LAGPeR to AG-ReID dataset.

Method	BB	$A \rightarrow G$		$G \rightarrow A$	
		Rank-1	mAP	Rank-1	mAP
BoT [17]	ViT	33.15	22.7	28.90	20.32
VDT [38]	ViT	34.74	23.42	29.83	21.53
SeCap(Ours)	ViT	**37.93**	**24.96**	**30.87**	**22.99**

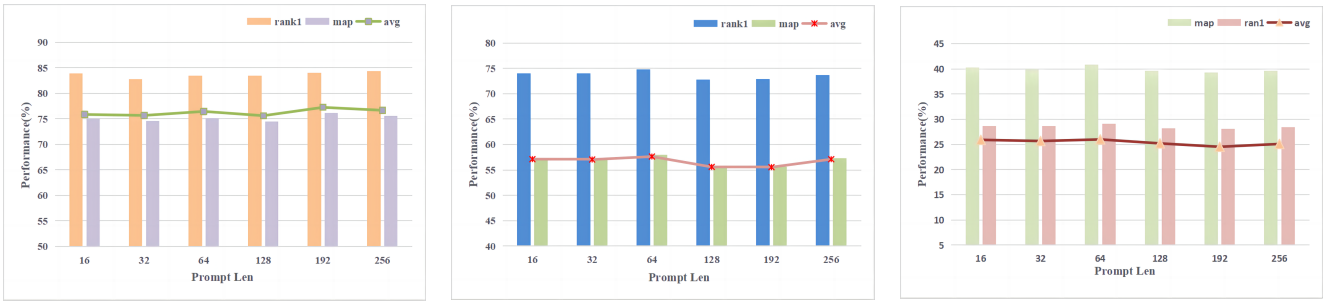
For the Prompt Re-calibration Module(PRM), we explore different methods of incorporating view-invariant features by comparing the Add, Cat, and Attn methods (#1 vs #2 vs #6). Add represents the method of integrating view-invariant features into the prompts through addition; Cat involves concatenating view-invariant features to the prompts and integrating them via self-attention; Attn involves learning view-invariant information through the attention mechanism, which is the method used in **PRM**. Among these methods, Attn achieves the best results.
For the LFRM module, we compare the effects of using two-way attention and Transformer decoding blocks (#3 vs #6). The two decoding structures are shown in Fig. 8, where the two-way attention(Two-Way) demonstrate significant performance improvements. Additionally, we validate the effectiveness of the feature fusion module (#4 vs #6), confirming its utility. Lastly, we assess the impact of overlapping patches (#5 vs #6), which also contribute to performance enhancement.

12. Parameter Analysis

As illustrated in Fig. 6, we analyze the impact of the hyperparameter λ on the model's performance. When λ is set to

(a) AG-ReID (b) G2APS-ReID (c) LAGPeR

Figure 6. Fig. 6a ∼ Fig. 6c show the impact of hyperparameter λ on model performance under three datasets. For simplicity, only setting $A \rightarrow G$ is shown on the AGPReID datasets. Rank1 and mAP are reported (%). The avg represents the average performance of mAP.



(a) AG-ReID (b) G2APS-ReID (c) LAGPeR

Figure 7. Fig. 7a ∼ Fig. 7c show the impact of prompt length L on model performance under three datasets. For simplicity, only setting $A \rightarrow G$ is shown on the AGPReID datasets. Rank1 and mAP are reported (%). The avg represents the average performance of mAP.

0.001, the SeCap model performs best on the G2APS-ReID and LAGPeR datasets. For the AG-ReID dataset, the optimal λ is 10. **This discrepancy arises because the G2APS-ReID and LAGPeR datasets have a higher number of IDs, necessitating a smaller coefficient to balance the difficulty between viewpoint classification and ID classification.**

Under the identical λ setting, we carry out a detailed analysis of the impact of prompt length L on the model's performance. As presented in Fig. 7, **the model's performance is not highly sensitive to the prompt length L.** The model attains the best performance when the prompt length is set to 64.

13. Broader impact

The proposed method can be applied to existing aerial-ground person re-identification tasks, aiming to improve the performance of AGPReID tasks. All experiments are based on publicly available datasets, reconstructed datasets from public datasets, and datasets from public datasets, with the core goal of optimizing the application effect of the recognition model in real-world scenarios, rather than deliberately designing privacy leakage mechanisms. However, it is necessary to be vigilant against potential negative effects, such as the privacy leakage risks that may arise from using surveillance and drone-captured person re-identification data. Therefore, when collecting such data, we ensure that relevant individuals are fully informed and strictly manage and use the data to protect individuals' privacy rights and interests.

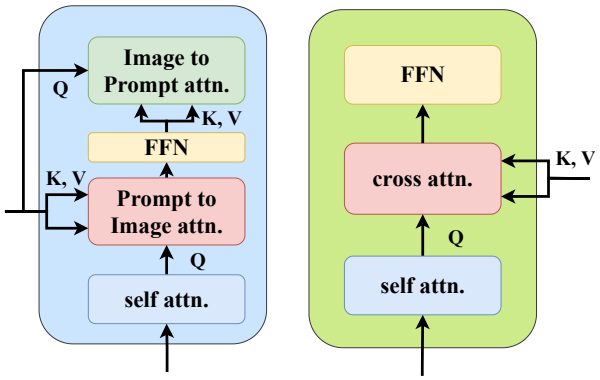


Figure 8. The structure of Two-Way attention and Transformer Decoding Block.

14. Scene examples of the LAGPeR dataset.

As Fig.9, we show some scene examples of the LAGPeR dataset, where A-Cam represents the aerial view camera and G-Cam represents the ground view camera. Each column represents the same scene.

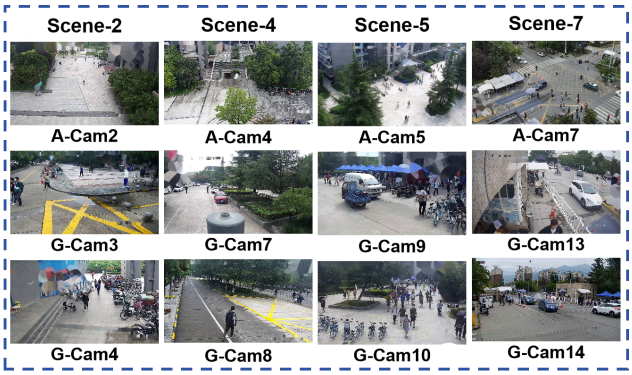


Figure 9. Example images from the LAGPeR, where each column represents images from the same scene.

15. Performance of $G \leftrightarrow G$

Our method achieves better results than top-performing compared methods in the $G \leftrightarrow G$ setting. We reconstructed the LAGPeR dataset and created the $G \leftrightarrow G$ subset to evaluate our model's performance, and the results are presented in Tab.10.

Table 10. Performance comparison on the $G \leftrightarrow G$ setting.

METHOD	ViT	VDT	TransReID	SeCap
Rank-1	78.36	77.18	79.23	**79.57**
mAP	77.38	76.35	78.09	**78.26**