Supplementary Material

Anonymous submission

A. Overview

In this supplementary material, we provide detailed dataset characteristics, additional experimental details and visualizations provided to complement the main paper. Specifically, the structure is organized as follows:

1. **Comprehensive Characteristics of CP2108:**
   - Image distributions in CP2108
   - Attribute annotations in CP2108
2. **Additional Experimental Details:**
   - Implementation Details
   - Extended comparison on AG-ReID.v2 and CARGO
   - Ablation study on key components of our SVPR-ReID
   - Hyperparameter sensitivity analysis of our SVPR-ReID
3. **Additional Visualizations:**
   - Attention heatmaps of global and local features
   - Ranklist comparison across different models

These analyses provide a comprehensive understanding of our SVPR-ReID and further validate the effectiveness of the proposed modules.

B. Comprehensive Characteristics of CP2108.

Image distributions

Image numbers. As shown in Fig. 1, the number of images from the UAV view is roughly twice that of the ground view. This is because UAV perspectives provide richer viewpoint diversity, while ground views are relatively static. Additionally, the day-to-night sample ratio is close to 1:1, ensuring a balanced and realistic distribution for evaluating cross-time and cross-platform person ReID performance.

Image pixels. The variations in camera viewpoints and distances introduce significant disparities in image resolutions across platforms,with resolutions ranging from 12 × 35 pixels to 1366 × 1440 pixels. As shown in Fig. 2, the UAV-captured images typically exhibit smaller body sizes, mostly concentrated around 90 pixels in height, due to higher-altitude perspectives. In contrast, the ground-based images primarily fall within the 150–250 pixel range, offering greater detail from closer viewpoints. These pronounced differences in image scale pose significant challenges for achieving robust AGPReID under severe resolution and appearance variations.

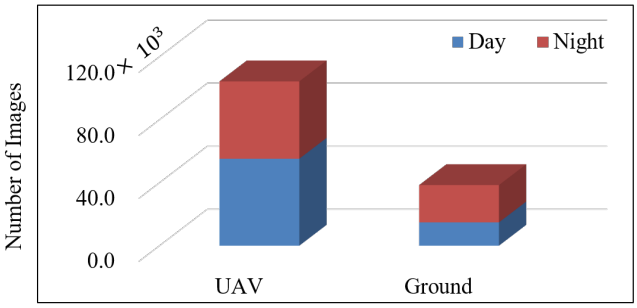


Figure 1: Distribution of image numbers in the CP2108 dataset.

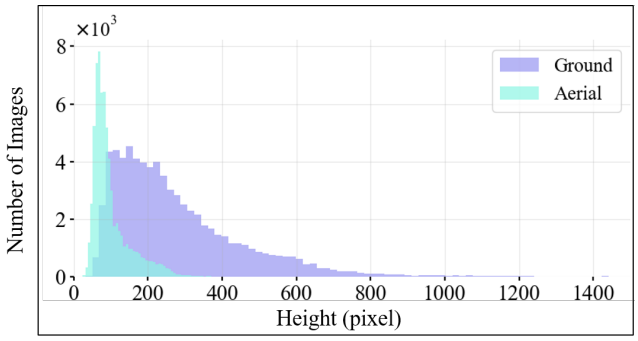
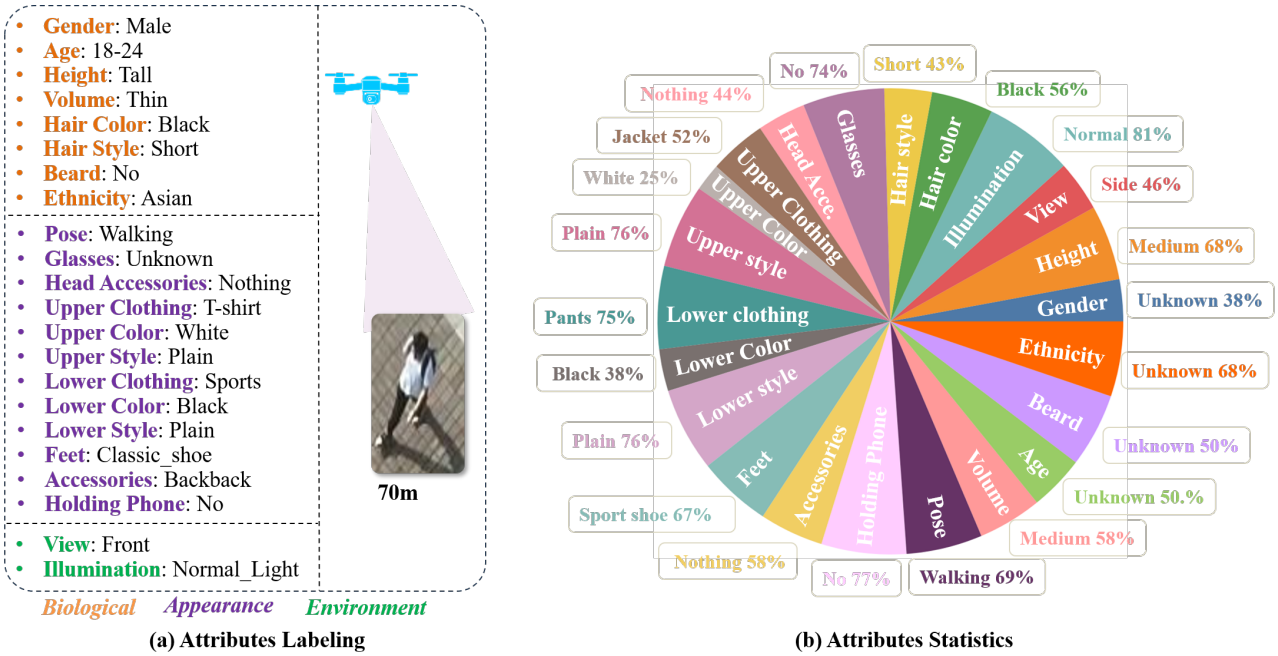


Figure 2: Distribution of the body heights (in pixels) across two platforms (aerial and ground) in the CP2108 dataset.

Attribute Annotations in CP2108

As shown in Fig. 3, we provide a detailed illustration of the attribute annotations on the CP2108 dataset.

First, as illustrated in Fig. 3 (a), we annotate 22 pedestrian attributes based on prior work, grouped into three categories: *Biological*, *Appearance*, and *Environment*. Notably, due to the frequent presence of mobile phones in real-world scenarios, we explicitly add a *Holding Phone* attribute, which captures an important behavioral cue for identity verification. Additionally, diverse pedestrian poses in CP2108, such



(a) Attributes Labeling **(b) Attributes Statistics**

Figure 3: Statistics of our CP2108 dataset including detailed attributes labels and statistics.

as walking, running, and riding, enrich contextual variation.

Second, Fig. 3 (b) presents the statistical distribution of attribute values, where some values occur more frequently (e.g., walking, plain style of clothes), while the overall distribution remains sufficiently diverse to reflect realistic variations. This realistic distribution ensures robust evaluation under various conditions.

Overall, these attributes serve as semantic priors to enhance AGPReID robustness against drastic viewpoint changes and image quality variations. Biological traits (e.g., gender, body volume) provide stable identity cues, while appearance-based attributes (e.g., clothing color, accessories) deliver fine-grained discriminative details. Environmental attributes such as view type and illumination further improve adaptability to platform-specific differences and challenging conditions like low light and backlighting. Integrating these semantics facilitates more effective disentanglement of viewpoint noise and strengthens identity representation consistency across diverse cross-platform scenarios.

C. Additional Experimental Details

Implementation Details

The experiments are conducted using PyTorch on a NVIDIA GTX 4090 GPU. We adopt the CLIP-Base-16 (Li, Sun, and Li 2023), pretrained on ImageNet, which is pre-trained on ImageNet. All AGPReID datasets are resized to 256×128 for both training and inference. The visual encoder operates with a patch size and stride of 16×16, and the token embedding dimension is set to $d = 768$. Data augmentation includes random horizontal flipping, random cropping, and random erasing. We train the model using the Adam optimizer with an initial learning rate of 5×10^{-5} for 120 epochs. Each mini-batch consists of 128 images sampled from 32 identities, with 4 instances per identity. More details are provided in the supplementary material, and the source code will be released upon acceptance.

Extended Comparisons

We further evaluate our method on two additional datasets: AG-ReID.v2 (Nguyen et al. 2024) and CARGO (Zhang et al. 2024), as shown in Table 1 and Table 2.

AG-ReID.v2 Dataset. As an extended version of AG-ReID.v1 (Nguyen et al. 2023), this dataset enhances diversity by introducing more identities and an additional wearable camera type, resulting in a more comprehensive and challenging setting for AGPReID evaluation. It contains 100,502 images from 1,605 identities captured by three types of cameras: CCTV, UAV, and wearable devices. The dataset is split into 807 identities for training and 798 for testing, and includes 15 attributes to facilitate cross-view matching. Evaluation is conducted under four protocols: $A \rightarrow C$, $C \rightarrow A$, $A \rightarrow W$, and $W \rightarrow A$, covering various view-transfer scenarios.

Performance on AG-ReID.v2. As shown in Table 1, single-view methods such as CLIP-ReID and TransReID achieve strong results on intra-view protocols ($A \rightarrow A$ and $C \rightarrow C$), reaching Rank-1 scores exceeding 90%. However, performance declines under cross-view protocols like $A \rightarrow W$ and $W \rightarrow A$, reflecting the challenge of aerial-ground variation. Among cross-view methods, AG-ReID.v2 achieves the highest Rank-1 accuracy of 93.62% in the $A \rightarrow W$ setting by explicitly leveraging head-centric features through its Elevated-View Attention Stream. Our method achieves

Method	ALL		A → C		C → A		A → W		W → A	
	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
Swin (Liu et al. 2021)	-	-	68.76	57.66	68.49	56.15	68.80	57.70	64.40	53.90
HRNet-18 (Wang et al. 2020)	-	-	75.21	65.07	76.26	66.17	76.25	66.16	76.25	66.17
SwinV2 (Liu et al. 2022)	-	-	76.44	66.09	80.08	69.09	77.11	62.14	74.53	65.61
MGN (Liu et al. 2021)	-	-	82.09	70.17	84.14	78.66	84.21	72.41	84.06	73.73
BoT (Luo et al. 2019)	-	-	85.40	77.03	84.65	75.90	85.17	73.06	84.65	75.90
ViT (Dosovitskiy et al. 2020)	-	-	85.40	77.03	84.65	75.90	85.65	75.90	84.27	76.59
TransReID (He et al. 2021)	72.75	63.11	85.98	78.93	81.12	72.57	88.22	80.04	84.72	72.96
CLIP-ReID* (Li, Sun, and Li 2023)	82.04	74.12	88.34	83.27	84.15	76.87	89.38	85.15	86.35	80.56
AGReID.v1 (Nguyen et al. 2023)	-	-	87.70	79.00	87.35	78.24	83.14	78.13	79.08	77.03
VDT (Zhang et al. 2024)	-	-	86.46	79.13	84.14	78.12	90.00	82.21	86.36	78.52
AGReID.v2 (Nguyen et al. 2024)	-	-	88.77	80.72	87.86	78.51	**93.62**	84.85	86.61	80.11
SeCap (Wang et al. 2025)	-	-	88.12	80.84	**88.24**	**79.99**	91.44	84.01	87.56	80.15
SVPR-ReID (Ours)	**82.29**	**74.38**	**89.25**	**84.35**	84.26	77.83	90.55	**85.71**	**87.78**	**80.64**

Table 1: Performance comparison with state-of-the-art methods under five testing protocols on AG-ReID.v2 dataset. "C", "W" and "A" denote ground view, wearable view and aerial view, respectively. The best performances are in bold. Rank1 and mAP are reported in %.

Method	ALL		A → A		G → G		A ↔ G	
	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
SBS (He et al. 2023)	50.32	43.09	67.50	49.73	72.31	62.99	31.25	29.00
PCB (Sun et al. 2018)	51.00	44.50	55.00	44.60	74.10	67.60	34.40	30.40
BoT (Luo et al. 2019)	54.81	46.49	65.00	49.79	77.68	66.47	36.25	32.56
MGN (Wang et al. 2018)	54.81	49.08	65.00	52.96	83.93	71.05	31.87	33.47
VV (Kuma et al. 2019)	45.83	38.84	67.50	49.73	72.31	62.99	31.25	29.00
AGW (Ye et al. 2021)	60.26	53.44	67.50	56.48	81.25	71.66	43.57	40.90
ViT (Dosovitskiy et al. 2020)	61.54	53.54	80.00	64.47	82.14	71.34	43.13	40.11
TransReID (He et al. 2021)	60.90	53.17	80.00	68.24	83.93	74.77	42.50	38.06
CLIP-ReID (Li, Sun, and Li 2023)	67.63	62.08	67.50	63.08	83.04	77.59	57.50	52.54
VDT (Zhang et al. 2024)	64.10	55.20	**82.50**	66.83	82.14	71.59	48.12	42.76
SeCaP (Wang et al. 2025)	68.59	60.19	80.00	68.08	86.61	75.42	**69.43**	58.94
SVPR-ReID (Ours)	**70.51**	**66.00**	79.50	**70.59**	**86.71**	**80.85**	68.75	**64.05**

Table 2: Performance comparison with state-of-the-art methods under four testing protocols on CARGO dataset. "A" and "G" denote aerial view and ground view, respectively. The best performances are in bold. Rank1 and mAP are reported in %.

90.55% Rank-1 and 89.44% mAP under the same protocol, demonstrating competitive retrieval performance through the integration of attribute-aware prompts and progressive local refinement. The smaller viewpoint gap in AG-ReID.v2 reduces the advantage of our viewpoint disentanglement module, but our framework still maintains robust performance across all protocols, particularly in mAP, which reflects overall ranking quality and robustness under occlusion or illumination changes.

CARGO Dataset. The CARGO dataset is a synthetic AG-PReID benchmark generated using tools like MakeHuman (Briceno and Paul 2018) and Unity3D (Patil and Alvares 2015). It contains 108,563 images of 5,000 unique identities captured by 13 cameras, including 8 ground and 5 aerial views. The training set consists of 51,451 images from 2,500 identities, and the remaining 51,024 images from 2,500 identities form the test set. Evaluation is conducted under four protocols: ALL, $A \rightarrow A$, $G \rightarrow G$, and $A \leftrightarrow G$, where the ALL protocol overall retrieval performance, while the remaining protocols focus on specific intra-view and cross-view matching cases.

Performance on CARGO Dataset. As shown in Table 2, single-view methods such as AGW and TransReID perform reasonably well on intra-view settings ($A \rightarrow A$ and $G \rightarrow G$), achieving Rank-1 scores above 80%. However, their performance drops drastically under the cross-view $A \leftrightarrow G$ setting, with Rank-1 accuracy below 60% for most methods, reflecting the challenge of bridging large appearance gaps. Among cross-view methods, SeCap achieves high performance with 69.43% Rank-1 and

58.94% mAP. Our method further improves these results to 68.75% in Rank-1 and 64.05% in mAP, yielding a 5.11% mAP gain, which better reflects overall retrieval quality. This improvement primarily stems from our framework's ability to jointly address viewpoint bias, incorporate global semantic guidance through attribute-aware prompts, and progressively refine local details, enabling robust and discriminative representations for challenging cross-view scenarios.

Additional Ablation study

Ablation on ASMoE. We examine the impact of different attribute-scattering strategies in ASMoE. As shown in Table 3, the MoE-based approach achieves the best performance with 59.23% Rank-1 and 42.40% in mAP under the ALL protocol. In contrast, MLP and Attention variants show lower results across both protocols. These findings confirm that MoE provides stronger adaptability and generalization for modeling diverse attribute semantics compared to fixed-structure alternatives.

Type of ASMoE	ALL		A ↔ G	
	Rank-1	mAP	Rank-1	mAP
MLP	57.58	41.76	63.94	46.64
Attention	56.31	40.89	61.79	46.53
ASMoE	**59.23**	**42.40**	**63.94**	**47.37**

Table 3: Comparison among attention, MLP and MoE designs for the ASMoE module under two evaluation protocols, reporting Rank-1 and mAP (%).

Ablation on Attribute Formats. We explore the effect of attribute organization formats in ASMoE by comparing full-sentence textual descriptions (Text) and word-level phrases (Word), as shown in Table 4. Textual descriptions are generated by an LLM that converts attribute labels into natural language sentences. The results indicate that both formats achieve comparable accuracy, with word-level phrases achieving slightly higher performance, 59.23% Rank-1 and 42.40% mAP under the ALL protocol. This suggests that compact phrases are more efficient and reduce redundancy compared to long descriptive text. Meanwhile, exploring more efficient attribute utilization strategies with LLMs remains a promising direction for future work.

Ablation on Attribute Groups. We conduct an ablation study under the ALL protocol using three primary categories (*appearance*, *biological*, and *environment*) and additionally isolate clothing-related attributes (*upper clothing* and *lower clothing*). Fig. 4 evaluates the contribution of different attribute groups. The appearance-related attributes achieve the highest performance, with 59.23% Rank-1 and 42.40% mAP, underscoring their strong discriminative capability for identity representation. This superiority can be attributed to clothing occupying a significant portion of the pedestrian region, providing rich fine-grained cues compared to other attribute types. While biological attributes offer complementary context, environment-related attributes contribute less fine-grained semantics, which may explain their relatively weaker performance.

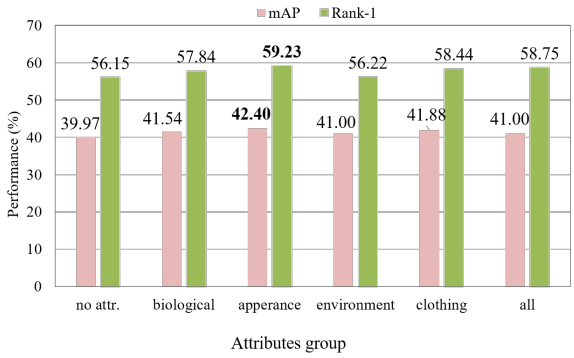


Figure 4: Ablation study on different attribute groups.

Attribute Type	ALL		A↔G	
	Rank-1	mAP	Rank-1	mAP
ALL-Text	57.47	41.10	63.14	45.68
ALL-Word	58.75	41.00	62.77	45.62
APP-Word	**59.23**	**42.40**	**63.94**	**47.37**

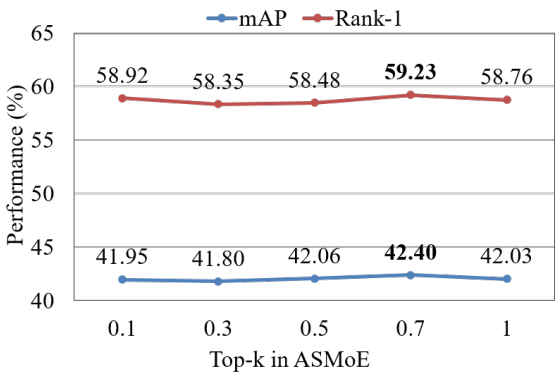
Table 4: Comparison of different attribute prompt designs in ASMoE. "ALL" denotes all annotated attributes, "APP" refers to appearance-related attributes, "Text" represents full-sentence descriptions, and "Word" indicates word-level phrases. Results are reported in Rank-1 accuracy and mAP (%), with the best results in bold.

Ablation on Selective Ratio of ASMoE. As shown in Fig. 5 (a), the Top-k ratio in ASMoE has a slight impact on overall performance. The best results are achieved when the ratio is set to 0.5, yielding 59.23% Rank-1 and 42.40% mAP. Extremely low or high ratios cause slight performance degradation, indicating that not all attributes remain effective under challenging conditions (e.g., occlusion, low illumination). A moderate selection helps focus on the most reliable attribute cues while avoiding noise.

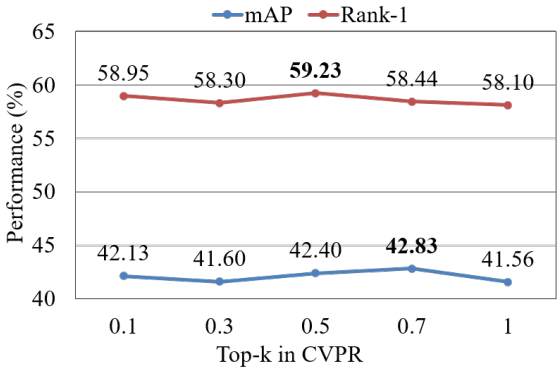
Ablation on Selective Ratio of CVPR. As shown in Fig. 5 (b), the performance trend is similar for CVPR. The highest Rank-1 is 58.44% and mAP is 42.83% at $k = 0.7$, while other settings yield comparable results. This indicates that selectively refining discriminative patches benefits representation quality, as some local regions may become unreliable under severe viewpoint or illumination variations.

Hyperparameter Analysis

Discussion of λ on view loss. As shown in Fig. 6 (a), performance remains relatively stable when λ ranges from 0.01 to 10, with the best result observed at $\lambda = 1$, achieving Rank-1 of 59.23% and mAP of 42.40%. This indicates that a balanced weight for view loss is beneficial for disentangling view-specific features without overwhelming the main ReID objective.



(a) Ablation of Top-k ratio in ASMoE



(b) Ablation of Top-k ratio in CVPR

Figure 5: Ablation of the selective ratio (Top-k) in SVPR-ReID.

Discussion of λ on attribute loss. As shown in Fig. 6 (b), we observe that attribute loss weight significantly influences performance. The model achieves the highest accuracy at $\lambda = 0.05$ with Rank-1 of 59.23% and mAP of 42.40%. Further increasing λ causes a gradual performance decline, indicating that excessive attribute supervision may dominate optimization and impair representation learning.

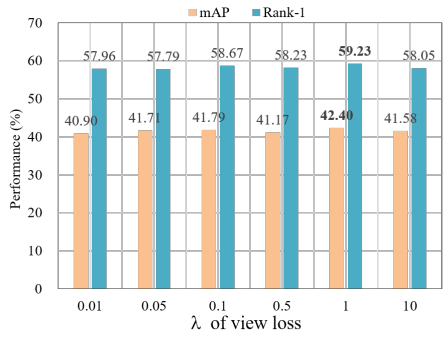
D. Visualization

Heatmaps

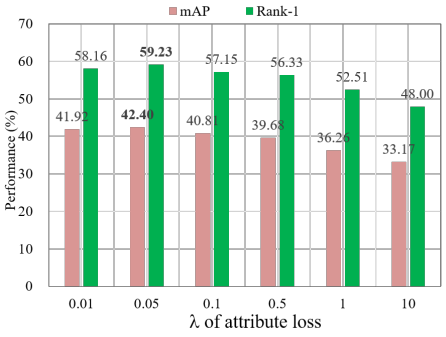
We visualize the attention maps of the SVPR-ReID and baseline models on global and local features. As shown in Fig. 7, there is some environmental noise in the baseline model. In contrast, our method focuses more accurately on key regions for identity matching, such as the head and clothing, under challenging conditions such as low illumination and extreme viewpoint changes.

Ranklist

Retrieval results on SVPR-ReID. As shown in Fig. 8, the visualization of the retrieval results shows that our SVPR-ReID focuses more on fine-grained identity cues compared to the baseline, highlighting the importance of incorporating detailed attribute semantics for robust performance in complex AGPReID scenarios.



(a) Ablation of different view weight



(b) Ablation of different attributes weight

Figure 6: Impact of hyperparameter λ on attribute loss and view loss under the ALL protocol.

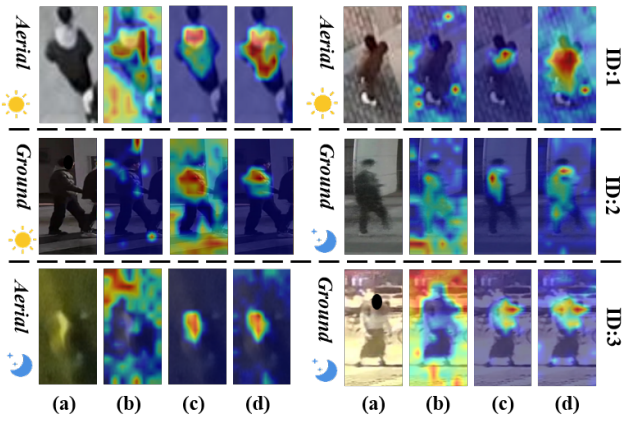


Figure 7: Visualization of attention maps for our SVPR-ReID framework. (a) Input images, (b) Baseline method, (c) Baseline + VBD and (d) SVPR-ReID (Ours).

Retrieval results on attribute features. As shown in Fig. 9, attribute labels align well with the model's attention on corresponding semantic regions, even under challenging cross-view conditions, showing the role of attribute-aware prompts in improving fine-grained alignment and retrieval robustness.

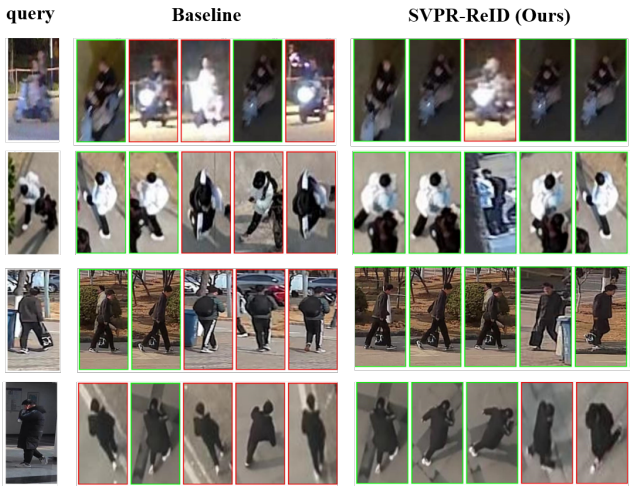


Figure 8: Comparison of several retrieval visualizations on the CP2108 dataset of protocol ALL. Red and green boxes represent wrong and correct matchings. Each row shows the top-5 ranked images for a given query (left) on different methods. Correct matches are marked in green and incorrect matches in red.

Figure 9: Visualizations of retrieval results based on attribute features.

References

Briceno, L.; and Paul, G. 2018. Makehuman: a review of the modelling framework. In *Proceedings of the Congress of the International Ergonomics Association*, 224–232. Springer.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

He, L.; Liao, X.; Liu, W.; Liu, X.; Cheng, P.; and Mei, T. 2023. Fastreid: A pytorch toolbox for general instance re-identification. In *Proceedings of the ACM International Conference on Multimedia*, 9664–9667.

He, S.; Luo, H.; Wang, P.; Wang, F.; Li, H.; and Jiang, W. 2021. Transreid: Transformer-based object re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, 15013–15022.

Kuma, R.; Weill, E.; Aghdasi, F.; and Sriram, P. 2019. Vehicle re-identification: an efficient baseline using triplet embedding. In *Proceedings of the International Joint Conference on Neural Networks*, 1–9. IEEE.

Li, S.; Sun, L.; and Li, Q. 2023. Clip-reid: exploiting vision-language model for image re-identification without concrete text labels. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, 1405–1413.

Liu, Z.; Hu, H.; Lin, Y.; Yao, Z.; Xie, Z.; Wei, Y.; Ning, J.; Cao, Y.; Zhang, Z.; Dong, L.; et al. 2022. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12009–12019.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

Luo, H.; Gu, Y.; Liao, X.; Lai, S.; and Jiang, W. 2019. Bag of Tricks and a Strong Baseline for Deep Person Re-Identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Nguyen, H.; Nguyen, K.; Sridharan, S.; and Fookes, C. 2023. Aerial-Ground Person Re-ID. arXiv:2303.08597.

Nguyen, H.; Nguyen, K.; Sridharan, S.; and Fookes, C. 2024. AG-ReID.v2: Bridging Aerial and Ground Views for Person Re-Identification. *IEEE Transactions on Information Forensics and Security*, 19: 2896–2908.

Patil, P. P.; and Alvares, R. 2015. Cross-platform application development using unity game engine. *Int. J*, 3(4).

Sun, Y.; Zheng, L.; Yang, Y.; Tian, Q.; and Wang, S. 2018. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European conference on computer vision*, 480–496.

Wang, G.; Yuan, Y.; Chen, X.; Li, J.; and Zhou, X. 2018. Learning discriminative features with multiple granularities for person re-identification. In *Proceedings of the ACM international conference on Multimedia*, 274–282.

Wang, J.; Sun, K.; Cheng, T.; Jiang, B.; Deng, C.; Zhao, Y.; Liu, D.; Mu, Y.; Tan, M.; Wang, X.; et al. 2020. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10): 3349–3364.

Wang, S.; Wang, Y.; Wu, R.; Jiao, B.; Wang, W.; and Wang, P. 2025. SeCap: Self-Calibrating and Adaptive Prompts for Cross-view Person Re-Identification in Aerial-Ground Networks. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 22119–22128.

Ye, M.; Shen, J.; Lin, G.; Xiang, T.; Shao, L.; and Hoi, S. C. 2021. Deep learning for person re-identification: A survey and outlook. *IEEE transactions on pattern analysis and machine intelligence*, 44: 2872–2893.

Zhang, Q.; Wang, L.; Patel, V. M.; Xie, X.; and Lai, J.-H. 2024. View-decoupled Transformer for Person Re-identification under Aerial-ground Camera Network. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22000–22009.