GPS: GEOMETRY-GUIDED PATCH SYNTHESIS FOR AERIAL-GROUND PERSON RE-IDENTIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Aerial-ground person re-identification must match identities despite substantial changes in local appearance and visibility across viewpoints. Existing methods primarily modify or align observed images, but do not explicitly construct identity-preserving cross-view variations with controllable matching difficulty. In this paper, we propose Geometry-Guided Patch Synthesis (GPS), a feature-level hard-positive learning framework that synthesizes structured changes in local identity evidence during training. Specifically, a Geometry-Constrained Patch Synthesizer (GCPS) module uses a fixed spatial coordinate prior to reorganize patch tokens through three complementary operations: compression, merging, and vanishing, thereby producing candidates with controlled local variations. Then, a Hard-Positive Patch Synthesis (HPPS) module is proposed to select the hardest identity-consistent candidate within a prescribed difficulty range relative to a real same-identity cross-view observation. Finally, a Dual-Path Evidence Fusion (DPEF) module transfers locally supported correspondence between the original and selected synthetic features into identity learning while preserving unsupported original evidence. Extensive experiments on four aerial-ground benchmarks demonstrate the effectiveness of GPS, with consistent improvements over Plain CLIP across all evaluated protocols. Component ablations further demonstrate the complementary contributions of structured patch synthesis, difficulty-controlled selection, and evidence fusion. The code will be made publicly available.

1 INTRODUCTION

Aerial-ground person re-identification (AG-ReID) retrieves the same person across aerial and ground cameras (Nguyen et al., 2023; 2024). It extends identity association in ground-camera networks (Ye et al., 2022; Luo et al., 2019) to observations with sharply different elevations and visible surfaces (Zhang et al., 2021a; Yan et al., 2021). Clothing patterns, carried objects, and visible body regions provide local identity cues (Sun et al., 2021; He et al., 2021). The task requires associating different appearances of the same person while retaining cues that distinguish identities.

Recent AG-ReID methods strengthen the encoding and alignment of observed images (Zhang et al., 2024b; Wang et al., 2025a; Li et al., 2025; Zhang et al., 2026; Yang & Zhang, 2026). Augmentation broadens training observations through local erasing and patch mixing (Zhong et al., 2020; Yun et al., 2019), but these edits can obscure discriminative details, introduce cues from other identities, or combine noncorresponding regions across views (Figure 1(a)). These transformations diversify visible content without explicitly modeling the coupled changes in local extent, overlap, and visibility induced by aerial-ground imaging. Hard-positive learning emphasizes difficult same-identity matches (Hermans et al., 2017; Yu et al., 2018), while synthesis varies their local evidence (Huang et al., 2018). Our premise is that geometry-guided synthesis, regulated by identity consistency, exposes the model to informative intra-identity variations and encourages discrimination when local cues change or disappear. When familiar details become mixed, weakened, or unavailable, learning to associate the altered observation with the same person encourages the model to draw on other informative cues. This extends the configurations of local appearance covered by identity supervision and reduces reliance on a fixed arrangement of local evidence. These observations motivate a complementary question: how can we construct challenging positives that reflect aerial-ground changes in the spatial extent, prominence, and visibility of local evidence, preserve identity, and regulate the additional difficulty relative to a real same-identity cross-view pair?

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

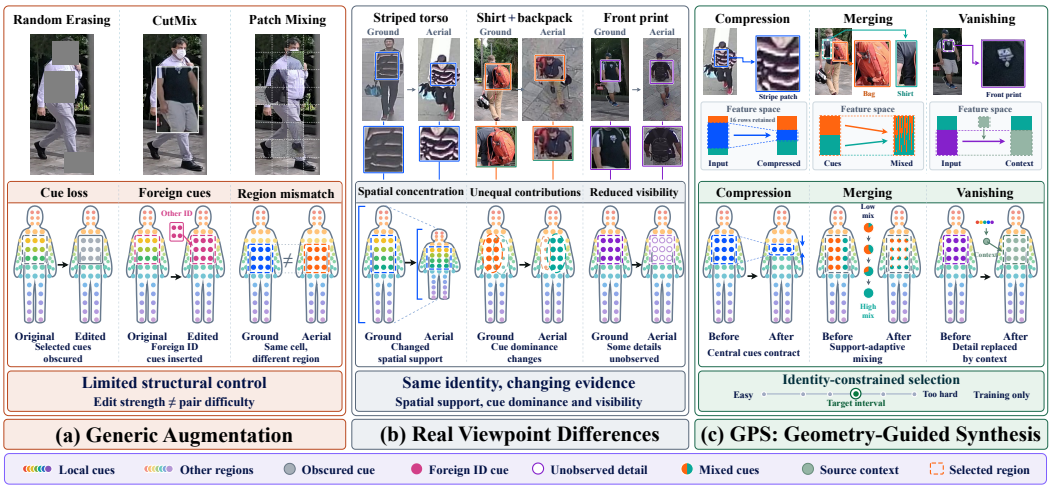


Figure 1: Motivation for GPS. (a) Generic local augmentation lacks explicit structural and pair-difficulty control. (b) Real same-identity AG-ReID v2 pairs illustrate changes in spatial support, cue prominence, and visibility. (c) GPS combines compression, merging, and vanishing in feature space with identity-constrained selection, retaining the output token grid. Photo crops provide qualitative references, while overlays and cue diagrams are schematic.

Aerial-ground geometry provides a useful starting point. Orientation and visibility determine which regions support identity matching (Sun & Zheng, 2019; Miao et al., 2019). A clothing pattern spread across several ground-view patches can occupy a much smaller region in an aerial image. Regions that are separated in the ground view can overlap in projection, changing which local cues contribute to the observation. Changes in visible surfaces also alter the relative prominence of clothing and carried objects, such as a shirt and a backpack, so these cues contribute unequally across views. Distinctive details can also become invisible, making other visible regions more important for identity matching (Wang et al., 2020; Huang et al., 2023) (Figure 1(b)). These changes motivate compression, merging, and vanishing as three complementary operations for positive-sample synthesis. They address situations in which local evidence becomes spatially concentrated, visible cues contribute unequally, and distinctive details become unavailable. Because the body projection changes, patches at the same image-grid position may also depict different regions (Zhang et al., 2026). Together, these effects produce difficult matches even when identity remains the same, especially when appearance cues relied on in ground images weaken. Compression and merging allow training to vary the organization and retention of local evidence, while vanishing introduces the additional challenge of recognizing an identity from the remaining cues. These edits diversify local evidence for identity learning across aerial and ground observations.

We propose Geometry-Guided Patch Synthesis (GPS). Its core Geometry-Constrained Patch Synthesizer (GCPS) uses a fixed coordinate prior to guide compression, merging, and vanishing in patch-token space. Compression contracts the longitudinal extent of local evidence through resampling, merging adjusts how much local evidence is retained according to relative support, and vanishing suppresses selected details through source-context replacement. GCPS combines these variations with observed same-identity cross-view information to generate candidates at different strengths. For this synthesis to support identity learning, the constructed challenge should be assessed together with the identity evidence retained by the candidate. Measuring difficulty relative to a real cross-view pair connects the additional challenge to the observed appearance variation of that identity. Hard-Positive Patch Synthesis (HPPS) then selects the hardest identity-consistent candidate within a prescribed range of additional difficulty relative to the observed cross-view pair (Figure 1(c)). As synthesis changes which local cues remain available, Dual-Path Evidence Fusion (DPEF) identifies supported correspondences between the original and selected synthetic features and integrates their shared evidence into identity learning while retaining the original observation as a reference. Together, these modules turn structured local variation into supervision for cross-view recognition. Synthesis and fusion operate during training, while retrieval uses a single image encoder and its global descriptor to rank gallery images by cosine distance.

The main contributions are summarized as follows:

- We propose GPS, a geometry-guided hard-positive learning framework for AG-ReID. It connects structured local variation with identity-constrained positive-sample construction, enriching the evidence configurations encountered during training and supporting recognition under changes in local appearance and visibility.
- We design GCPS to synthesize compression, merging, and vanishing in patch-token space with a fixed coordinate prior. HPPS controls difficulty relative to real cross-view pairs, while DPEF integrates supported correspondences into identity supervision.
- Experiments on AG-ReID v1, AG-ReID v2, CARGO, and LAGPeR show consistent gains over Plain CLIP, including an absolute 9.88% Rank-1 improvement on LAGPeR under $G \rightarrow AG$. Component ablations and evaluations across backbones and ReID frameworks demonstrate complementary benefits and broad applicability.

2 Related Work

Recent AG-ReID methods improve cross-view representations through feature decomposition, local adaptation, and semantic alignment. VDT separates view-related and view-unrelated features (Zhang et al., 2024b). GSAlign combines geometric compensation with visibility-consistent alignment (Li et al., 2025), and HiHR organizes identity consistency and view-specific detail hierarchically (Yang & Zhang, 2026). Part-based aggregation and patch interactions preserve fine-grained identity cues (Sun et al., 2021; He et al., 2021), while SeCap and ViSA adapt local representations and correspondences to viewpoint changes (Wang et al., 2025a; Zhang et al., 2026). These studies advance the organization and association of local identity evidence across views.

Augmentation and synthesis enrich training observations through local edits and controlled generation. Random erasing and CutMix remove or exchange local content (Zhong et al., 2020; Yun et al., 2019), while VIF-AGReID introduces spatially varying patch rotation and mixing (Khalid et al., 2025). Conditional generation further expands pose and viewpoint coverage (Grolleau et al., 2026; Wang et al., 2026). Hard-positive learning selects challenging same-identity associations or constructs additional difficult observations (Hermans et al., 2017; Yu et al., 2018; Huang et al., 2018). GPS combines geometry-guided patch-feature synthesis with pair-relative difficulty selection and identity constraints, using compression, merging, and vanishing to construct structured local variations. Further discussion appears in Appendix D.

3 Methodology

3.1 Overview and Formulation

GPS constructs local-evidence variations for cross-view identity learning. Geometry-Constrained Patch Synthesizer (GCPS) generates candidates through compression (C), merging (M), and vanishing (V), collectively denoted CMV. Hard-Positive Patch Synthesis (HPPS) selects their difficulty, and Dual-Path Evidence Fusion (DPEF) fuses shared original–synthetic evidence (Figure 2).

For an image x with identity y and platform ν, a shared CLIP encoder (Radford et al., 2021) produces patch tokens $P \in \mathbb{R}^{N \times D}$ and a class token $c \in \mathbb{R}^{D}$. We denote source and same-identity opposite-platform donor observations by (P^o, c^o) and (P^d, c^d). Either platform can be the source. The least similar eligible cross-view observation supplies the donor, anchoring candidate construction to a challenging real pair. The source retains identity supervision throughout training, while selected synthetic observations add local and global signals. Appendix B details the formulations and parameter settings for CMV, candidate construction, HPPS, DPEF, and the training objectives.

3.2 Geometry-Constrained Patch Synthesizer

The aerial-ground gap alters the distribution, relative contribution, and visibility of local cues. GCPS models these changes with three complementary operations, as illustrated in Figure 1(c).

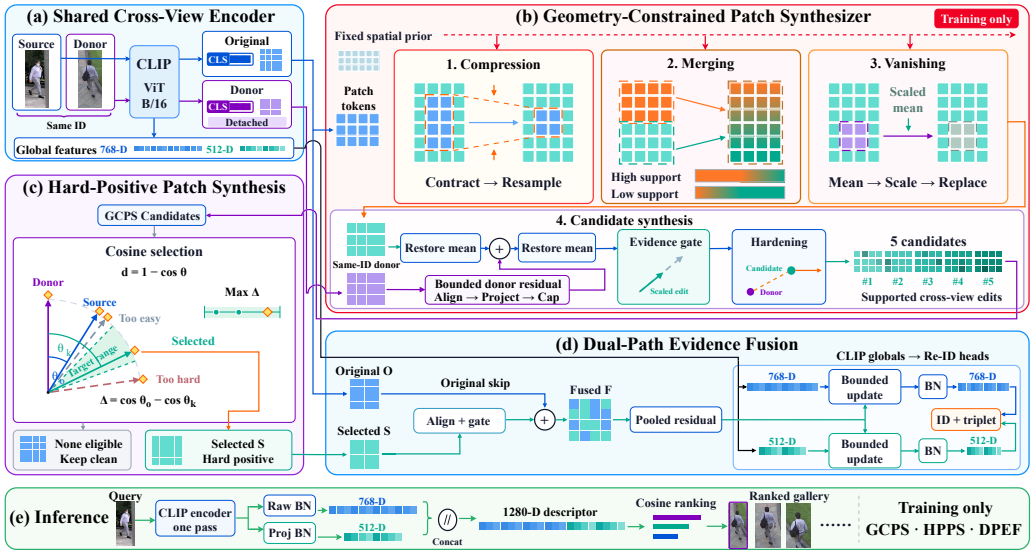


Figure 2: Overview of GPS. (a) A shared encoder supplies source and same-identity cross-view donor features. (b) GCPS constructs candidates. (c) HPPS selects a challenging positive. (d) DPEF integrates shared and path-specific evidence for training. (e) Inference uses a single image encoding. Token grids and gallery order are schematic.

Fixed geometric guidance. Each token receives a fixed prior $\xi_i = (u_i, v_i, z_i)$: vertical and horizontal coordinates plus positional support z_i. GCPS, DPEF, and the topology objective share the encoder's image grid. At strength s, the anisotropic neighborhood used for patch merging is:

$$A_{ij}(s) = \mathrm{softmax}_j\left(-\frac{1}{2}\left[\frac{(u_i-u_j)^2}{\sigma_u(s)^2} + \frac{(v_i-v_j)^2}{\sigma_v(s)^2}\right]\right). \tag{1}$$

The bandwidths $\sigma_u(s)$ and $\sigma_v(s)$ control direction-dependent aggregation. A broader vertical neighborhood lets evidence interact across more body-height positions, whereas the narrower horizontal extent limits lateral mixing. This directional structure governs local evidence sharing in merging and preserves a common spatial reference for subsequent operations. Below, $A = A(s)$.

Patch compression. An elevated viewpoint can concentrate spatially separated cues into a smaller projected area. We model this effect by contracting a central longitudinal interval and resampling its evidence onto the fixed token grid. Let $g_s : [0,1] \to [0,1]$ be a piecewise-linear forward coordinate map with input knots $0, a, b, 1$, where $a = 0.2$, $b = 0.8$, and $\delta_s = \frac{1}{2}\alpha_c(s)(b-a)$ with $0 \le \alpha_c(s) < 1$. We define the map and compressed tokens by

$$\big(g_s(0), g_s(a), g_s(b), g_s(1)\big) = (0, a+\delta_s, b-\delta_s, 1), \quad p_i^c = \mathrm{Interp}\big(P^o; g_s^{-1}(u_i), v_i\big). \tag{2}$$

Here, g_s maps source to output coordinates, and Interp performs channel-wise linear interpolation between neighboring source rows in the same column. The central interval shrinks by a factor of $1-\alpha_c(s)$, while the upper and lower margins expand to preserve the endpoints. The map remains strictly increasing, so inverse sampling preserves vertical order and requires no padding. Increasing $\alpha_c(s)$ concentrates the central evidence over a smaller longitudinal span. Resampling produces $P^c \in \mathbb{R}^{N\times D}$ on the fixed output coordinates ξ_i, providing the input to the subsequent merging operation. Compression thus changes the spatial extent of local evidence, and merging subsequently adjusts its relative contribution according to each token's support within its neighborhood.

Patch merging. Projection overlap changes which cues dominate a local observation. Merging captures this asymmetry through relative support. Let e_i^c be the min–max normalized norm of p_i^c and combine feature response with positional support as $\psi_i = \omega z_i + (1-\omega)e_i^c$, where $\omega \in [0,1]$:

$$\beta_i = \mathrm{sigmoid}\left(\kappa\,\underbrace{[\psi_i - (A\psi)_i]}_{\text{relative support}}\right), \qquad p_i^m = p_i^c + \underbrace{\alpha_m(s)(1-\beta_i)}_{\text{adaptive strength}}\big[(AP^c)_i - p_i^c\big]. \tag{3}$$

The slope κ sets support sensitivity and $\alpha_m(s)$ controls mixing strength. A strongly supported token has larger β_i and retains more of its own evidence, while a weaker token receives more neighborhood evidence. Merging thus refines compression according to local prominence. Even within the same geometric neighborhood, tokens can undergo different amounts of mixing. This distinction matters when overlapping regions contribute unequally: the operation lets prominent evidence remain identifiable while less supported evidence moves toward the surrounding context.

Patch vanishing. Viewpoint changes can weaken or remove local identity cues. Vanishing suppresses selected details so that learning also draws on the remaining evidence. We recompute normalized responses e_i^m and define risk $r_i^v = \omega(1 - z_i) + (1 - \omega)(1 - e_i^m)$. The mask m_i^v selects the highest-risk fraction ρ_v. With source context $\mu_o = N^{-1} \sum_i p_i^o$, we replace these tokens by:

$$p_i^v = \underbrace{(1 - m_i^v) p_i^m}_{\text{retained evidence}} + \underbrace{m_i^v \eta_v \mu_o}_{\text{context replacement}} . \tag{4}$$

The context scale η_v controls the replacement. Subsequent mean restoration preserves the source context statistic before donor residual synthesis. Compression, merging, and vanishing therefore alter the spatial extent of evidence, how strongly it contributes, and whether its local detail remains available. Because risk is recomputed after merging, vanishing responds to the evidence remaining after earlier edits. Its selected locations can therefore differ from those most affected by compression or merging, allowing complementary changes within a candidate.

Cross-view candidates. The donor anchors geometric edits to observed same-identity variation. After aligning donor evidence to the source, we normalize local residuals relative to the source and project them away from its class-token direction. Bounded residuals enter supported locations with large cross-view differences. This combines the spatial redistribution induced by the operators with appearance changes observed in the donor. The evidence gate then modulates the combined edit, while donor-relative tangent hardening adjusts candidate difficulty as training progresses. Soft alignment, bounded donor residuals, and an evidence gate yield the transformation $\mathcal{T}$, producing candidates $\mathcal{C} = \{(P^k, c^k) = \mathcal{T}_{\ell_k}(P^o, c^o; P^d, c^d)\}_{k=1}^K$ at strength levels ℓ_k. A training curriculum increases their variation. Candidate construction also includes sparse distractor injection. HPPS evaluates the resulting matching difficulty relative to the observed source–donor pair.

3.3 Learning from Synthesized Hard Positives

Identity-constrained hard-positive selection. An edit's difficulty depends on the image and current encoder. HPPS therefore compares candidates against a real cross-view donor. Let h_o, h_d, h_k combine normalized class-token and pooled patch information, and let $d(a, b) = 1 - \cos(a, b)$. The identity-consistent set $\mathcal{I}$ requires source proximity, identity-prediction agreement, and sufficient confidence. HPPS chooses the hardest eligible candidate within a target interval:

$$\Delta_k = \underbrace{d(h_k, h_d)}_{\text{candidate–donor distance}} - \underbrace{d(h_o, h_d)}_{\text{source–donor distance}} , \qquad k^\star = \underset{\substack{k \in \mathcal{I} \\ \delta_{\min} \le \Delta_k \le \delta_{\max}}}{\arg\max} \Delta_k . \tag{5}$$

Positive Δ_k measures the additional challenge relative to the observed pair. The lower bound encourages a meaningful challenge, while the upper bound limits its extent. The interval regulates difficulty while identity constraints retain the source as the learning target. Using the observed source–donor pair as the reference separates the extra challenge introduced by synthesis from the difficulty already present in the pair. Thus, selection responds to how much an edit changes cross-view matching for that identity. As the encoder improves, stronger candidates can replace previously challenging ones under the same identity constraints. When no raw candidate meets the difficulty interval and an over-hard candidate is available, HPPS can calibrate it toward the source under the same acceptance conditions. Samples without an accepted candidate retain their clean representation.

Dual-path evidence fusion. DPEF compares original tokens $O = P^o$ with selected synthetic tokens $S = P^{k^\star}$ to identify evidence that persists through the variation. Semantic similarity, coordinate proximity, and positional support align synthetic evidence into $\widetilde{S}$ and produce support scores w_i. We normalize support by each position's geometric ceiling U_i to select common evidence:

$$w_i^{\text{rel}} = \frac{w_i}{\max(U_i, \epsilon)}, \qquad m_i^c = \mathbf{1}\big[w_i^{\text{rel}} \ge \theta_c\big] . \tag{6}$$

Here, ϵ is a numerical floor. Relative support governs common-evidence membership, while raw w_i retains geometric weighting in fusion and the objectives. Noncommon positions with normalized original response at least as strong as the normalized aligned synthetic response form mask m_i^o. The remainder are synthetic-only. The fused token is:

$$o_i^f = o_i + \tanh(\alpha)\left[\underbrace{m_i^c w_i W(\widetilde{s}_i - o_i)}_{\text{shared-evidence update}} - \underbrace{\eta_o m_i^o (o_i - \text{mean}(O))}_{\text{original-detail adjustment}}\right]. \tag{7}$$

Shared evidence supplies a weighted residual through projection W, while original-only details receive a small context-centered adjustment scaled by η_o. Synthetic-only evidence supplies no residual. The learned gate $\tanh(\alpha)$ controls fusion, preserving the original path as the reference. This division connects the training signal to the intended synthesis: the shared state highlights cues supported through the change, while the original-only state preserves access to details present in the real observation. Keeping support selection separate from update strength also gives distinct roles to semantic agreement and geometric reliability. A location can qualify as common evidence yet receive a modest update when its absolute support is low. The pooled residual updates the raw and projected global identity streams through a norm-controlled interface.

3.4 Optimization and Inference

The base objective $\mathcal{L}_{\text{base}}$ combines identity and triplet learning, cross-view alignment, contrastive learning, hard-negative ranking, and teacher-guided synthetic supervision. Geometry-Valid Patch Synthesis (GVPS) encourages coherent edits with sustained identity support. Original-Synthetic Identity Consistency (OSIC) aligns global identities and supported local evidence while maintaining donor-relative difficulty. The total objective is:

$$\mathcal{L} = \mathcal{L}_{\text{base}} + \lambda_{\text{G}}\mathcal{L}_{\text{GVPS}} + \lambda_{\text{O}}\mathcal{L}_{\text{OSIC}}. \tag{8}$$

At inference, each image is encoded once. Raw and projected BN-neck features are concatenated and ranked by cosine distance. Synthesis and evidence learning are used only during training.

4 Experiments

We evaluate GPS on four aerial–ground benchmarks, examine its components, and relate retrieval improvements to learning from synthesized hard positives. See Appendices A and C for details.

4.1 Datasets and Evaluation Protocols

We use the official splits of AG-ReID v1 (Nguyen et al., 2023), AG-ReID v2 (Nguyen et al., 2024), CARGO (Zhang et al., 2024b), and LAGPeR (Wang et al., 2025a), covering bidirectional aerial–ground, within-view, and mixed-gallery retrieval. CARGO supplies synthetic imagery and serves as the primary ablation benchmark, while the other three use captured images. We report mAP and CMC Rank-1 (R1) in percentages under each benchmark's camera exclusions and valid-query rules. A, G, C, and W denote aerial, ground, CCTV, and wearable cameras, respectively. G→AG matches ground queries against a mixed gallery containing both aerial and ground images.

4.2 Implementation Details

The primary model uses CLIP ViT-B/16 (Radford et al., 2021) with 256×128 inputs and a clean 1280-dimensional retrieval descriptor. Matching uses cosine distance without test-time augmentation or reranking. Plain CLIP uses its native identity and triplet objectives. Module ablations use a matched clean baseline to assess candidate synthesis, difficulty selection, and evidence fusion.

4.3 Comparison with Existing Methods

GPS improves on Plain CLIP in every evaluated protocol (Table 1). It achieves the highest mAP in all six AG-ReID v1/v2 directions and leads both metrics in all three LAGPeR protocols. On CARGO, it leads the ALL and A→G scores among the listed methods. Its absolute 9.72% mAP improvement

Table 1: Performance comparison on AG-ReID v1/v2, LAGPeR, and CARGO (%). Best and second-best results are bold and underlined.

Method \ Dataset Protocol	(a) AG-ReID v1				(b) AG-ReID v2							
	A→G		G→A		A→C		C→A		A→W		W→A	
	mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1
VDT (Zhang et al., 2024b)	74.44	82.91	78.57	86.59	79.13	86.46	78.12	86.14	82.21	90.00	78.52	85.26
SeCap (Wang et al., 2025a)	76.16	84.03	78.34	87.01	80.84	88.12	79.99	88.24	84.01	91.44	80.15	87.56
GSAlign (Li et al., 2025)	75.01	83.75	77.73	84.10	81.38	87.86	81.05	88.02	83.98	90.63	80.90	87.31
SD-ReID (Wang et al., 2026)	75.40	**85.16**	77.02	85.97	80.61	87.04	79.24	86.74	84.06	90.86	80.12	86.79
ViSA (Zhang et al., 2026)	75.56	84.91	77.94	84.10	83.61	89.43	82.32	88.57	85.99	91.63	83.10	89.23
SAS-VPReID (Yang et al., 2026)	75.98	84.24	77.28	86.17	76.39	83.40	78.47	85.26	82.10	89.45	78.47	85.73
Plain CLIP	72.53	79.93	75.46	83.16	77.02	83.83	75.33	83.66	79.87	86.92	75.96	83.03
GPS	**77.95**	84.69	**81.34**	**88.25**	**84.75**	**89.90**	**83.58**	**89.01**	**88.29**	**93.35**	**85.11**	**90.98**

Method \ Dataset Protocol	(c) LAGPeR						(d) CARGO							
	A→G		G→A		G→AG		ALL		A→G		A→A		G→G	
	mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1
VDT (Zhang et al., 2024b)	28.97	40.15	31.98	33.55	16.45	19.50	55.20	64.10	42.76	48.12	66.83	**82.50**	71.59	82.14
SeCap (Wang et al., 2025a)	30.37	41.79	33.42	35.26	19.24	24.39	60.19	68.59	58.94	69.43	68.08	80.00	75.42	86.61
GSAlign (Li et al., 2025)	27.96	38.74	31.00	32.01	19.41	24.52	57.95	65.06	61.55	64.89	65.55	80.00	73.86	83.04
SD-ReID (Wang et al., 2026)	29.72	40.15	32.86	34.47	18.88	22.85	57.47	65.06	46.44	53.12	67.70	**82.50**	74.08	81.25
ViSA (Zhang et al., 2026)	30.42	41.37	33.59	35.06	20.31	25.61	65.46	70.51	69.00	71.28	67.78	**82.50**	**83.90**	88.39
SAS-VPReID (Yang et al., 2026)	30.76	43.83	33.40	35.42	23.02	31.12	58.48	64.42	57.36	58.51	63.41	70.00	76.90	83.93
Plain CLIP	28.78	40.74	31.91	33.13	20.69	26.30	63.49	69.55	63.63	65.96	66.36	75.00	79.08	85.71
GPS	**35.67**	**48.56**	**39.02**	**41.63**	**27.85**	**36.18**	**71.55**	**77.24**	**73.35**	**74.47**	**73.24**	80.00	83.87	**90.18**

Table 2: Component ablations on CARGO (mAP and Rank-1, %). C, M, and V denote compression, merging, and vanishing. ✓/blank entries denote enabled/disabled components.

(a) Compression, merging, and vanishing

C	M	V	ALL		A→G	
			mAP	R1	mAP	R1
			64.39	69.55	65.44	65.96
✓			67.88	74.04	68.80	70.21
	✓		67.72	73.08	69.56	69.15
		✓	66.81	74.68	66.72	67.96
✓	✓		69.41	75.64	70.65	72.34
✓		✓	69.63	75.00	70.83	70.21
	✓	✓	69.36	75.32	71.46	72.34
✓	✓	✓	**71.55**	**77.24**	**73.35**	**74.47**

(b) HPPS and DPEF modules

GCPS	HPPS	DPEF	ALL		A→G	
			mAP	R1	mAP	R1
			64.39	69.55	65.44	65.96
✓			70.05	75.96	71.65	72.34
✓	✓		71.21	76.92	72.35	73.40
✓		✓	70.99	75.96	72.12	73.40
✓	✓	✓	**71.55**	**77.24**	**73.35**	**74.47**

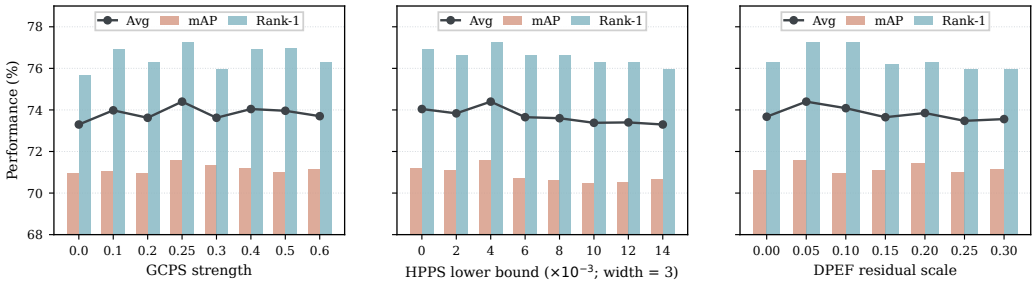
(c) GVPS and OSIC objectives

GVPS	OSIC	ALL		A→G	
		mAP	R1	mAP	R1
		69.63	75.00	70.83	70.21
✓		70.39	75.00	71.31	72.34
	✓	70.79	76.60	71.24	71.28
✓	✓	**71.55**	**77.24**	**73.35**	**74.47**

under A→G exceeds the within-view gains, consistent with learning to recognize identities when local evidence is reorganized or weakened. Improvements across synthetic and captured images support applicability beyond a particular image source, while within-view gains indicate preserved identity discrimination when queries and gallery images share the same viewing platform.

4.4 Compression, Merging, and Vanishing Ablation

Within the shared donor-based synthesis path, each operator improves retrieval (Table 2(a)). Merging yields the strongest single-operator cross-view mAP, and vanishing gives the highest single-operator ALL Rank-1. Every pair improves mAP over its constituent operators on both protocols,



(a) The effect of s (GCPS) (b) The effect of δ_{min} (HPPS) (c) The effect of ρ (DPEF)

Figure 3: Sensitivity to (a) GCPS strength, (b) the HPPS difficulty lower bound, and (c) DPEF residual scale. The interval width in (b) is 0.003. Defaults are 0.25, 0.004, and 0.10, respectively. Bars show mAP and Rank-1, and the line shows their average.

while combining all three gives the best results in every column. These complementary gains support learning changes in cue organization, relative contribution, and visibility together.

4.5 Module and Loss Ablation

GCPS, HPPS, and DPEF. GCPS provides the largest gain over the matched clean baseline (Table 2(b)). HPPS and DPEF each further improve mAP on both protocols, and their combination performs best. These results support combining synthesis with identity-consistent difficulty selection and supported-evidence fusion. Each module includes its dependent supervision.

GVPS and OSIC. With the architecture fixed, GVPS and OSIC each improve mAP, and their combination gives the best results in every column, with larger joint gains for cross-view retrieval (Table 2(c)). The objectives complement coherent, identity-supported edits with global and local consistency between original and altered observations of the same identity.

4.6 Parameter Sensitivity

Figure 3 evaluates eight settings each for GCPS and HPPS and seven for DPEF. The three sweeps vary the strength coefficient used to construct GCPS candidates, the lower bound of the HPPS difficulty interval with its width held fixed, and the coefficient of the pooled DPEF residual in the global feature update. Moderate synthesis strength and difficulty yield the strongest combined performance, while DPEF benefits from a modest residual contribution. These trends support controlling difficulty and evidence strength throughout hard-positive learning.

4.7 Visual Analysis

Figure 4(a,b) compares the Plain CLIP initialization with GPS under identical retrieval protocols. Correct identities move ahead of visually similar distractors in the displayed examples. For all 94 valid A→G queries against the full ground gallery, GPS increases the mean and median nearest-match margins and improves Rank-1 and Rank-5 accuracy. The margin $m(q) = d^-(q) - d^+(q)$ compares cosine distances to the nearest incorrect and correct identities. A positive value places a correct identity first. These observations connect the retrieval gains to better separation of matching identities from competing appearances. They are consistent with training on hard positives that challenge local appearance correspondence while maintaining identity consistency.

Figure 4(c–h) further compares the cross-view feature distributions of ViT, Plain CLIP, and GPS on CARGO. GPS increases the distance gap between same-identity and different-identity pairs to $\Delta = 0.614$, compared with 0.442 for Plain CLIP and 0.323 for ViT. The t-SNE projections also show tighter grouping of aerial and ground samples from the same identity and clearer separation among the visualized identities. Together with the retrieval analysis, these results support learning identity-discriminative features that remain consistent under cross-view changes in local appearance.

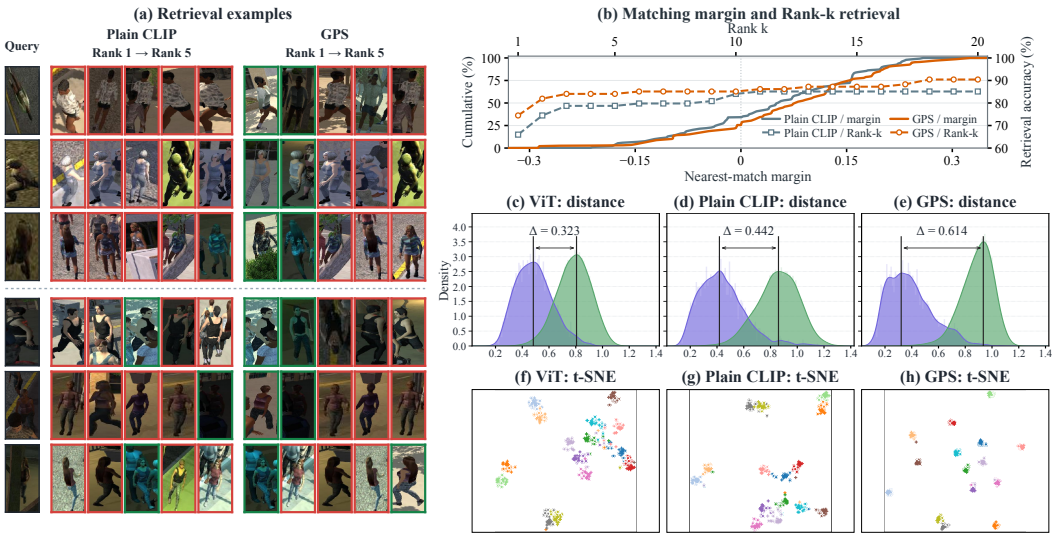


Figure 4: CARGO visual analysis. (a) Top-five retrieval: A→G (upper three rows), ALL (lower three). Green/red borders mark correct/incorrect matches. (b) Margin CDFs (solid, bottom/left axes, monotone interpolation) and Rank-k accuracy (dashed with markers, top/right axes). (c–e) Cross-view same/different-identity cosine-distance distributions, with Δ marking the gap. (f–h) t-SNE: colors denote identities, and marker shapes distinguish aerial and ground samples.

Table 3: GPS across backbones and ReID frameworks on CARGO ALL (mAP and Rank-1, %). Δ denotes percentage-point gains computed before rounding.

Base model	Base		+ GPS		Δ	
	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
Plain CLIP (Radford et al., 2021)	63.49	69.55	71.55	77.24	+8.07	+7.69
ViT (Dosovitskiy et al., 2021)	53.54	61.54	57.06	65.38	+3.52	+3.84
VDT (Zhang et al., 2024b)	55.20	64.10	59.24	66.71	+4.04	+2.61
SeCap (Wang et al., 2025a)	60.19	68.59	68.34	72.12	+8.15	+3.53
DINOv3 (Siméoni et al., 2025)	65.25	69.32	69.07	77.56	+3.82	+8.24
TransReID (He et al., 2021)	51.59	60.58	53.24	62.82	+1.66	+2.24

4.8 Generality across Backbones and ReID Frameworks

Paired experiments with Plain CLIP, ViT, DINOv3, VDT, SeCap, and TransReID show improvements in both retrieval metrics under the CARGO ALL protocol (Table 3). Operating on patch-feature sequences lets GPS complement different pretrained representations and native ReID objectives. These comparisons extend hard-positive learning beyond the primary CLIP configuration: local synthesis can supply additional identity-learning signals while each model retains its native feature representation and the retrieval heads used by its baseline configuration.

5 Conclusion

GPS treats aerial–ground viewpoint change as a redistribution of local identity evidence. GCPS constructs candidates through compression, merging, and vanishing. HPPS selects them by relative difficulty and identity support, and DPEF routes aligned original–synthetic evidence into global identity learning. The complementary gains from compression, merging, and vanishing support learning changes in local cue organization, contribution, and visibility together, while selection and fusion improve how these synthesized positives contribute to identity learning. The auxiliary construction–selection–fusion path shapes the clean image descriptor during training and is removed at inference, retaining the encoder and cosine-distance retrieval. Experiments on four benchmarks show consistent improvements over Plain CLIP.

AI USE STATEMENT

AI assistants supported manuscript organization, reference summarization, inspection of supplied code and experiment records, bilingual drafting and editing, and LaTeX preparation. The authors are responsible for the final technical content and references.

ETHICS STATEMENT

Person re-identification uses information about appearance and movement. The evaluation uses established research benchmarks under their data access and usage terms. Applications involving identifiable people require appropriate consent, access controls, and privacy protections.

REPRODUCIBILITY STATEMENT

Section 3 specifies the synthesis, selection, fusion, and training objectives. Section 4 defines the benchmark comparisons, ablation studies, and analysis protocols. Appendix B details the coordinate definitions, loss weights, and training configuration.

REFERENCES

Shuoyi Chen, Mang Ye, and Bo Du. Rotation invariant transformer for recognizing object in UAVs. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 2565–2574, 2022. doi: 10.1145/3503161.3547799. URL https://doi.org/10.1145/3503161.3547799.

Yifei Deng, Chenglong Li, Yuyang Zhang, Guyue Hu, and Jin Tang. Cross-modal fuzzy alignment network for text-aerial person retrieval and a large-scale benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 38732–38741, 2026.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Yingying Feng, Jie Li, Jie Hu, Yukang Zhang, Lei Tan, and Jiayi Ji. MDReID: Modality-decoupled learning for any-to-any multi-modal object re-identification. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-1418.

William Grolleau, Astrid Sabourin, Guillaume Lapouge, and Catherine Achard. 3D-LENS: A 3D lifting-based elevated novel-view synthesis method for single-view aerial-ground re-identification, 2026. URL https://arxiv.org/abs/2604.26520. Accepted at ECCV 2026.

Ruiyang Ha, Songyi Jiang, Bin Li, Bikang Pan, Yihang Zhu, Junjie Zhang, Xiatian Zhu, Shaogang Gong, and Jingya Wang. Multi-modal multi-platform person re-identification: Benchmark and method. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10251–10261, 2025.

Lingxiao He, Xingyu Liao, Wu Liu, Xinchen Liu, Peng Cheng, and Tao Mei. FastReID: A Pytorch toolbox for general instance re-identification. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 9664–9667, 2023. doi: 10.1145/3581783.3613460.

Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. TransReID: Transformer-based object re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15013–15022, 2021.

Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *CoRR*, abs/1703.07737, 2017. URL https://arxiv.org/abs/1703.07737.

Zhangyi Hu, Bin Yang, and Mang Ye. Empowering visible-infrared person re-identification with large foundation models. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-3726.

Houjing Huang, Dangwei Li, Zhang Zhang, Xiaotang Chen, and Kaiqi Huang. Adversarially occluded samples for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5098–5107, 2018. doi: 10.1109/CVPR.2018.00535. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Huang_Adversarially_Occluded_Samples_CVPR_2018_paper.html.

Meiyan Huang, Chunping Hou, Qingyuan Yang, and Zhipeng Wang. Reasoning and tuning: Graph attention network for occluded person re-identification. *IEEE Transactions on Image Processing*, 32:1568–1582, 2023. doi: 10.1109/TIP.2023.3247159. URL https://doi.org/10.1109/TIP.2023.3247159.

Yuli Huang, Hongxu Chen, Zhanxiang Feng, and Jianhuang Lai. Perspective driven prototype alignment for aerial-ground person re-identification. In *Image and Graphics (ICIG 2025)*, volume 16162 of *Lecture Notes in Computer Science*, pp. 517–528. Springer Nature Singapore, 2026. doi: 10.1007/978-981-95-3393-0_42. URL https://link.springer.com/chapter/10.1007/978-981-95-3393-0_42.

Wajahat Khalid, Bin Liu, Xulin Li, Muhammad Waqas, and Muhammad Sher Afgan. Bridging the sky and ground: Towards view-invariant feature learning for aerial-ground person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9749–9758, 2025.

Jiachen Li and Xiaojin Gong. Prototypical contrastive learning-based CLIP fine-tuning for object re-identification. *arXiv preprint arXiv:2310.17218*, 2023. URL https://arxiv.org/abs/2310.17218.

Qiao Li, Jie Li, Yukang Zhang, Lei Tan, Jing Chen, and Jiayi Ji. GSAlign: Geometric and semantic alignment network for aerial-ground person re-identification. In *Advances in Neural Information Processing Systems*, volume 38, pp. 93785–93805, 2025. doi: 10.52202/085713-3137.

Siyuan Li, Li Sun, and Qingli Li. 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, pp. 1405–1413, 2023. doi: 10.1609/aaai.v37i1.25225.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022, 2021.

Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. 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 Workshops*, 2019.

Dongxu Mao, Shangzhi Teng, and Xueqiang Lyu. CVAF: A CLIP-based view-consistent alignment framework for aerial-ground person re-identification. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 22(3):85:1–85:19, 2026. doi: 10.1145/3785482. URL https://dl.acm.org/doi/10.1145/3785482.

Jiaxu Miao, Yu Wu, Ping Liu, Yuhang Ding, and Yi Yang. Pose-guided feature alignment for occluded person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 542–551, 2019.

Huy Nguyen, Kien Nguyen, Sridha Sridharan, and Clinton Fookes. Aerial-ground person re-id. In *IEEE International Conference on Multimedia and Expo*, pp. 2585–2590, 2023.

Huy Nguyen, Kien Nguyen, Sridha Sridharan, and Clinton Fookes. AG-ReID.v2: Bridging aerial and ground views for person re-identification. *IEEE Transactions on Information Forensics and Security*, 19:2896–2908, 2024.

Xuelin Qian, Yanwei Fu, Tao Xiang, Wenxuan Wang, Jie Qiu, Yang Wu, Yu-Gang Jiang, and Xiangyang Xue. Pose-normalized image generation for person re-identification. In *Proceedings of the European Conference on Computer Vision*, pp. 650–667, 2018.

Junyang Qiu, Zhanxiang Feng, Lei Wang, and Jianhuang Lai. Salient part-aligned and keypoint disentangling transformer for person re-identification in aerial imagery. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, 2024. doi: 10.1109/ICME57554.2024.10687753. URL https://doi.org/10.1109/ICME57554.2024.10687753.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, volume 139, pp. 8748–8763. PMLR, 2021.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025. URL https://arxiv.org/abs/2508.10104.

Xiaoxiao Sun and Liang Zheng. Dissecting person re-identification from the viewpoint of viewpoint. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 608–617, 2019.

Yifan Sun, Liang Zheng, Yali Li, Yi Yang, Qi Tian, and Shengjin Wang. Learning part-based convolutional features for person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(3):902–917, 2021. doi: 10.1109/TPAMI.2019.2938523.

Lei Tan, Yukang Zhang, Keke Han, Pingyang Dai, Yan Zhang, Yongjian Wu, and Rongrong Ji. RLE: A unified perspective of data augmentation for cross-spectral re-identification. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-4032.

Guan'an Wang, Shuo Yang, Huanyu Liu, Zhicheng Wang, Yang Yang, Shuliang Wang, Gang Yu, Erjin Zhou, and Jian Sun. High-order information matters: Learning relation and topology for occluded person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6449–6458, 2020. doi: 10.1109/CVPR42600.2020.00648. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Wang_High-Order_Information_Matters_Learning_Relation_and_Topology_for_Occluded_Person_CVPR_2020_paper.html.

Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning discriminative features with multiple granularities for person re-identification. In *Proceedings of the 26th ACM International Conference on Multimedia*, pp. 274–282, 2018. doi: 10.1145/3240508.3240552.

Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep high-resolution representation learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10):3349–3364, 2021. doi: 10.1109/TPAMI.2020.2983686.

Lei Wang, Quan Zhang, Junyang Qiu, and Jianhuang Lai. Rotation exploration transformer for aerial person re-identification. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, 2024a. doi: 10.1109/ICME57554.2024.10687588. URL https://doi.org/10.1109/ICME57554.2024.10687588.

Shining Wang, Yunlong Wang, Ruiqi Wu, Bingliang Jiao, Wenxuan Wang, and Peng Wang. SeCap: Self-calibrating and adaptive prompts for cross-view person re-identification in aerial-ground networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22119–22128, 2025a.

Yingquan Wang, Pingping Zhang, Dong Wang, and Huchuan Lu. Other tokens matter: Exploring global and local features of Vision Transformers for object re-identification. *Computer Vision and Image Understanding*, 244:104030, 2024b. doi: 10.1016/j.cviu.2024.104030. URL https://www.sciencedirect.com/science/article/pii/S1077314224001115.

Yuhai Wang and Maryam Pishgar. Dynamic token selective transformer for aerial-ground person re-identification. In *IEEE International Conference on Multimedia and Expo*, pp. 1–6, 2025. doi: 10.1109/ICME59968.2025.11210054.

Yuhao Wang, Pingping Zhang, Xuehu Liu, Zhengzheng Tu, and Huchuan Lu. Unity is strength: Unifying convolutional and transformeral features for better person re-identification. *IEEE Transactions on Intelligent Transportation Systems*, 26(3):3713–3723, 2025b. doi: 10.1109/TITS.2024.3521974.

Yuhao Wang, Xiang Hu, Lixin Wang, Pingping Zhang, and Huchuan Lu. SD-ReID: View-aware stable diffusion for aerial-ground person re-identification. *IEEE Transactions on Image Processing*, 35:5686–5697, 2026. doi: 10.1109/TIP.2026.3694135.

Ruiqi Wu, Bingliang Jiao, Wenxuan Wang, Meng Liu, and Peng Wang. Enhancing visible-infrared person re-identification with modality- and instance-aware visual prompt learning. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, pp. 579–588, 2024. doi: 10.1145/3652583.3658109.

Cheng Yan, Guansong Pang, Lei Wang, Jile Jiao, Xuetao Feng, Chunhua Shen, and Jingjing Li. BV-Person: A large-scale dataset for bird-view person re-identification. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10923–10932, 2021. doi: 10.1109/ICCV48922.2021.01076. URL https://doi.org/10.1109/ICCV48922.2021.01076.

Qiwei Yang and Pingping Zhang. HiHR: Hierarchical hyperbolic representation for aerial-ground person re-identification, 2026. URL https://arxiv.org/abs/2607.09186. Accepted at ECCV 2026.

Qiwei Yang, Pingping Zhang, Yuhao Wang, and Zijing Gong. SAS-VPReID: A scale-adaptive framework with shape priors for video-based person re-identification at extreme far distances. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops*, pp. 1599–1608, 2026.

Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven C. H. Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):2872–2893, 2022. doi: 10.1109/TPAMI.2021.3054775.

Rui Yu, Zhiyong Dou, Song Bai, Zhaoxiang Zhang, Yongchao Xu, and Xiang Bai. Hard-aware point-to-set deep metric for person re-identification. In *Computer Vision – ECCV 2018*, volume 11220 of *Lecture Notes in Computer Science*, pp. 196–212. Springer, 2018. doi: 10.1007/978-3-030-01270-0_12. URL https://link.springer.com/chapter/10.1007/978-3-030-01270-0_12.

Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. CutMix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6023–6032, 2019.

Pingping Zhang, Yuhao Wang, Yang Liu, Zhengzheng Tu, and Huchuan Lu. Magic tokens: Select diverse tokens for multi-modal object re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17117–17126, 2024a. doi: 10.1109/CVPR52733.2024.01620. URL https://openaccess.thecvf.com/content/CVPR2024/html/Zhang_Magic_Tokens_Select_Diverse_Tokens_for_Multi-modal_Object_Re-Identification_CVPR_2024_paper.html.

Pingping Zhang, Xiang Hu, Yuhao Wang, and Huchuan Lu. LATex: Leveraging attribute-based text knowledge for aerial-ground person re-identification, 2025. URL https://arxiv.org/abs/2503.23722.

Quan Zhang, Lei Wang, Vishal M. Patel, Xiaohua Xie, and Jianhuang Lai. View-decoupled transformer for person re-identification under aerial-ground camera network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22000–22009, 2024b.

Quan Zhang, Zeqiang Cai, Peiming Zhao, Jingze Wu, Cailun Wu, Hongbo Chen, and Jianhuang Lai. View-aware semantic alignment for aerial-ground person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4383–4392, 2026.

Shizhou Zhang, Qi Zhang, Yifei Yang, Xing Wei, Peng Wang, Bingliang Jiao, and Yanning Zhang. Person re-identification in aerial imagery. *IEEE Transactions on Multimedia*, 23:281–291, 2021a. doi: 10.1109/TMM.2020.2977528. URL https://doi.org/10.1109/TMM.2020.2977528.

Tianyu Zhang, Lingxi Xie, Longhui Wei, Zijie Zhuang, Yongfei Zhang, Bo Li, and Qi Tian. UnrealPerson: An adaptive pipeline towards costless person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11506–11515, 2021b. doi: 10.1109/CVPR46437.2021.01134. URL https://openaccess.thecvf.com/content/CVPR2021/html/Zhang_UnrealPerson_An_Adaptive_Pipeline_Towards_Costless_Person_Re-Identification_CVPR_2021_paper.html.

Yuxuan Zhao, Zhongao Zhou, Bin Yang, He Li, Jian Liang, Jun Chen, Bo Du, and Mang Ye. WHU-MARS: A multispectral aerial-ground benchmark towards any-scenario person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25461–25471, 2026.

Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 13001–13008, 2020. doi: 10.1609/aaai.v34i07.7000.

Ronghong Zhu, Hongxu Chen, Xiaohua Xie, and Jianhuang Lai. Global-local prompts-driven semantic guidance for aerial-ground person re-identification. In *Pattern Recognition and Computer Vision (PRCV 2025)*, volume 16287 of *Lecture Notes in Computer Science*, pp. 99–112. Springer Nature Singapore, 2026. doi: 10.1007/978-981-95-5755-4_7. URL https://link.springer.com/chapter/10.1007/978-981-95-5755-4_7.

A Additional Visualizations and Backbone Settings

A.1 Additional Visualizations

Retrieval on AG-ReID v2. In Figure 5, GPS corrects three top-ranked matches and increases another query's correct top-five matches from three to five. The last query's first correct match moves from fifth to second, leaving rank one unresolved. These examples complement Figure 4(a). Plain CLIP is the initialization reference, and both models share queries, galleries, and camera-exclusion rules.



Figure 5: Top-five retrieval comparisons on AG-ReID v2 under the A→C protocol. Each row shows an aerial query and the results of Plain CLIP and GPS from the same CCTV gallery. Green and red borders indicate correct and incorrect identities, respectively.

GCPS operation examples. Figure 6 illustrates compression, merging, and vanishing on three CARGO source images. Colored annotations show local longitudinal contraction, unequal contributions of local cues, and weakened local evidence. The final column combines these effects.

A.2 Paired Settings across Backbones and ReID Frameworks

Each Base/+ GPS pair matches the model variant, pretrained checkpoint, method-stage initialization, native prediction and retrieval heads, data split, preprocessing, sampler, training budget, optimizer, and checkpoint-selection rule. VDT, SeCap, and TransReID retain their native training objectives in both configurations. GPS operates on each model's patch-feature sequence, and evaluation uses the native clean descriptor path. Paired score differences measure the benefit of adding GPS within each model.

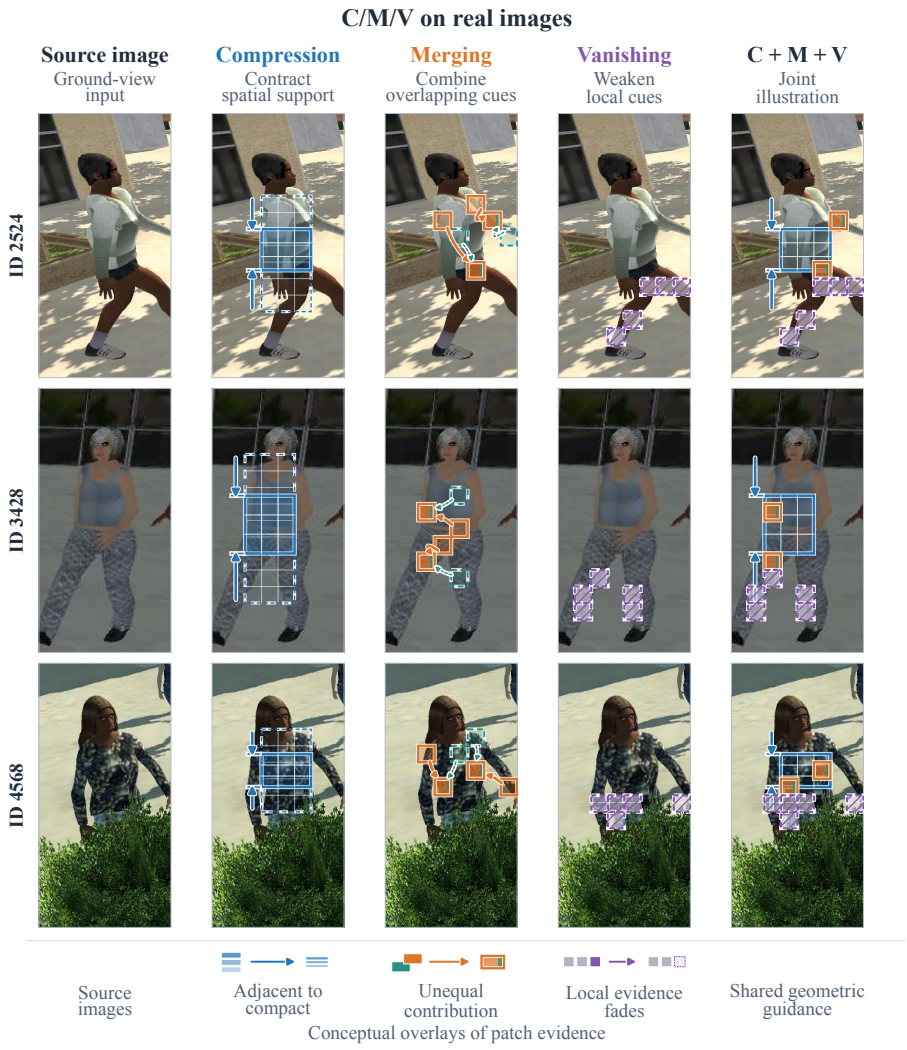


Figure 6: Illustration of GCPS operations on CARGO source images. Compression contracts a central longitudinal interval with compensating expansion at the margins. Merging combines local cues with unequal contributions, and vanishing attenuates local evidence. The final column illustrates their combined roles. Colored annotations depict the patch-level mechanisms.

B Method and Implementation Details

The CLIP ViT-B/16 image tower maps a 256×128 input to $N = 128$ patch tokens and a class token of dimension $D = 768$, together with a projected global feature of dimension $D' = 512$. The clean BN-neck outputs of the raw and projected identity heads form the 1280-dimensional retrieval descriptor.

B.1 GCPS Parameterization

GCPS and DPEF share fixed coordinates on the encoder's output token grid. For H_p rows and W_p columns, with $H_p, W_p > 1$ and $N = H_p W_p$, the zero-based row-major index i has coordinates

$$u_i = \frac{\lfloor i/W_p \rfloor}{H_p - 1}, \qquad v_i = 2\frac{i \bmod W_p}{W_p - 1} - 1,$$
$$z_i = \mathrm{clip}\left(1 - 0.25|v_i| - 0.15|u_i - 0.35|, 0, 1\right), \qquad \xi_i = (u_i, v_i, z_i). \tag{9}$$

The 256×128 input gives $H_p = 16$, $W_p = 8$, with vertical coordinate u_i and horizontal coordinate v_i. Other patch encoders use their own grid dimensions. The coordinate prior is fixed. GCPS correspondence combines cosine affinity at temperature 0.12 with half the Gaussian coordinate log-affinity at bandwidth 0.25. Its evidence gate uses semantic temperature 0.07 and coordinate bandwidth 0.25.

Spatial operators. For candidate level $\ell_k \in [0.35, 1]$ and training progress p, Equations 1–4 use strength $s_k = 0.25q(p)\ell_k$, where $q(p) = \mathrm{clip}(2p, 0, 1)$. Writing $s = s_k$,

$$\begin{aligned} \sigma_u(s) &= 0.10 + 0.32s, & \sigma_v(s) &= 0.08 + 0.12s, \\ \alpha_c(s) &= 0.10 + 0.55s, & \alpha_m(s) &= 0.05 + 0.55s. \end{aligned} \tag{10}$$

$\alpha_c(s)$ controls the central-interval contraction in Equation 2, $\alpha_m(s)$ controls support-adaptive mixing, and $\sigma_u(s)$ and $\sigma_v(s)$ set the merging neighborhood. The support mixture uses $\omega = 0.65$, merge slope $\kappa = 4$, and vanish context scale $\eta_v = 0.5$. For sequence X, the normalized response is $e_i(X) = (\|x_i\| - \min_j \|x_j\|)/\max(\max_j \|x_j\| - \min_j \|x_j\|, 10^{-6})$, evaluated on P^c for merging and recomputed on P^m for vanishing. The vanish mask selects the $n_v = \max(1, \min(N - 1, \mathrm{round}(N\rho_v)))$ highest-risk positions, with $\rho_v = 0.05\ell_k$. Mean restoration then preserves the source context:

$$\overline{P}^v = P^v - \mathbf{1}\,\mathrm{mean}(P^v) + \mathbf{1}\mu_o. \tag{11}$$

Here, $\mathbf{1}$ broadcasts a row vector across token positions.

Donor residuals. The least similar eligible same-identity cross-view observation serves as donor, with current-batch donors taking priority over the detached queue. We select 0.75 of eligible sources with the largest cross-view cosine gaps. Donor features and correspondence decisions are detached, while source features and selected candidates retain gradients.

After mean restoration, GCPS aligns donor tokens using soft semantic and coordinate correspondence, centers the aligned tokens, and rescales the donor residual to the edited source residual norm. From their difference, it removes the component parallel to the source class-token direction and caps the norm at $0.25\,\|c^o\|$. A sparse set of supported, high-gap tokens receives this residual before source-mean restoration. A learned gate initialized at zero modulates the combined spatial and donor candidate using source-candidate semantic and coordinate agreement, before HPPS hardening.

A distractor branch injects bounded different-identity residuals into 10% of tokens in 15% of source rows. Candidates retain source labels and undergo the same HPPS identity tests. Mixing coefficients appear in Table 4.

B.2 HPPS Selection

HPPS evaluates $K = 5$ GCPS candidates at evenly spaced scales in $[0.35, 1]$. Equation 5 uses $\delta_{\min} = 0.004$ and $\delta_{\max} = 0.007$. For numerically valid candidates $\mathcal{V}$, the identity-consistent set is

$$\begin{aligned} \mathcal{I} = \big\{ k \in \mathcal{V} : d(h_k, h_o) \le \epsilon_h, \quad p_y^k / p_y^o \ge \gamma_r, \\ p_y^k \ge \gamma_a, \quad \arg\max_j p_j^k = y \big\}. \end{aligned} \tag{12}$$

Thresholds are $\epsilon_h = 0.08$, $\gamma_r = 0.70$, and $\gamma_a = 0.10$. The same 768-dimensional raw-feature identity head yields p_y^o and p_y^k from source and candidate class tokens. Candidate evaluation uses running batch-normalization statistics. The absolute floor and predicted-class condition apply to the candidate.

Source identity distance and donor-relative difficulty use $h(P, c) = \widehat{c + q_P(P)}$, where $\widehat{v} = v/\|v\|$. With $\mu = N^{-1}\sum_i p_i$ and $a_i = \mathrm{softmax}_i(\|p_i - \mu\|/0.12)$, the focused pool is

$$q_P(P) = \|\mu\| \frac{0.65\mu + 0.35\sum_i a_i p_i}{\|0.65\mu + 0.35\sum_i a_i p_i\|}. \tag{13}$$

Numerical denominators are clamped away from zero.

Before selection, tokens and class features receive a negative donor-tangent residual, with coefficient at most 0.05 and direction

$$t = -\frac{\widehat{c}_d - (\widehat{c}_o^\top \widehat{c}_d)\widehat{c}_o}{\|\widehat{c}_d - (\widehat{c}_o^\top \widehat{c}_d)\widehat{c}_o\|}. \tag{14}$$

The vectors $\widehat{c}_o$ and $\widehat{c}_d$ are unit source and donor class features. We set $t = 0$ if the numerator vanishes. Token residuals scale with original norm, normalized edit magnitude, candidate level, and curriculum, while class-token residuals use the same factors except edit magnitude. At zero gate initialization, edit saliency starts from one tenth of normalized source-donor contrast and transitions continuously to the live edit magnitude as the gate grows, preserving distinct candidate strengths.

Let $\bar{n}_k = N^{-1}\sum_i \|p_i^k\|$ and $\bar{n}_o = N^{-1}\sum_i \|p_i^o\|$. Membership in $\mathcal{V}$ requires finite entries of P^k, $\bar{n}_k > 10^{-6}$, $0.25 \le \bar{n}_k/\bar{n}_o \le 4$, and absolute donor distance $d_k = d(h_k, h_d)$ within $[0, 1]$.

When no raw candidate meets the difficulty interval and at least one exceeds its upper bound, detached binary search targets $\Delta = 0.0055$ along $P(\lambda) = P^o + \lambda(P^k - P^o)$, $\lambda \in [0, 1]$, with the same interpolation for the class token. The calibrated candidate must belong to $\mathcal{I}$ and satisfy Equation 5.

B.3 DPEF Correspondence and Fusion

DPEF aligns the original and selected synthetic sequences using semantic similarity, coordinate proximity, and positional support. The coordinate term provides a soft locality prior on the fixed output grid:

$$M_{ij} = \mathrm{softmax}_j\left(\frac{\cos(o_i, s_j)}{\tau} - \frac{\|\xi_i - \xi_j\|^2}{2\sigma^2} + \log\sqrt{z_i z_j}\right), \qquad \widetilde{S} = MS. \tag{15}$$

Here, $\tau = 0.07$, $\sigma = 0.25$, and ξ_i follows Equation 9. Original token i has support

$$w_i = \max_j \left[\underbrace{\frac{1 + \cos(o_i, s_j)}{2}}_{\text{semantic agreement}} \underbrace{\exp\left(-\frac{\|\xi_i - \xi_j\|^2}{2\sigma^2}\right)\sqrt{z_i z_j}}_{\text{geometry and positional support}}\right]. \tag{16}$$

Writing the geometric and positional factor as G_{ij} gives the precomputed support ceiling

$$U_i = \max_j G_{ij}, \qquad 0 \le w_i \le U_i. \tag{17}$$

Equation 6 compares relative support against $\theta_c = 0.82$, with $\epsilon = 10^{-6}$. On the 16×8 grid at this bandwidth, $U_i = z_i$, so identical same-position features attain relative support one. Raw support w_i retains the geometric and positional weighting in fusion, GVPS, and OSIC.

Matching, raw support, and common-evidence decisions are detached, while aligned synthetic tokens remain differentiable. Among noncommon positions, DPEF compares original and aligned synthetic token norms, each normalized by its sequence's maximum: a greater or equal original response assigns the original-only mask, and the remainder is synthetic-only. Equation 7 uses $\eta_o = 0.08$, with W initialized to the identity and gate α to zero.

Global updates. Each accepted source receives a same-identity global mixture. For raw or projected source feature g_o and detached donor g_d,

$$\begin{aligned} v &= (1 - 0.10q)g_o + 0.10q\,\|g_o\|\,\widehat{g}_d, \qquad v' = \|g_o\|\,\widehat{v}, \\ g_b &= g_o + \min\left(1, \frac{0.04q}{\max(1 - \cos(g_o, v'), 10^{-6})}\right)(v' - g_o). \end{aligned} \tag{18}$$

Here, $q = \mathrm{clip}(2p, 0, 1)$. DPEF supplies $\delta = \mathrm{mean}(O^f - O)$ to the raw stream and $\delta\Pi$ to the projected stream, where $\Pi \in \mathbb{R}^{768\times512}$ is the CLIP projection. For either stream, set $v = g_b + 0.10\delta$, rescale it to $\|g_b\|$, and contract its displacement from g_o by $\min(1, 0.04/\max(1 - \cos(g_o, v), 10^{-6}))$. An exactly zero DPEF residual returns g_b.

Selected tokens also provide pooled identity-head supervision. Only the clean pass updates batch-normalization running statistics, while synthetic head passes preserve them.

B.4 Training Objective and Settings

GVPS and OSIC act at complementary stages of the learning problem. GVPS couples neighboring edits and their identity support, shaping the variations presented to the model. OSIC relates the resulting observation to its original identity at global and local levels. Their combination connects the structure of a synthetic positive with the identity relationships that its representation should preserve.

The native objective sums cross-entropy and hard-mined soft-margin triplet losses on both global heads. $\mathcal{L}_a$ is the mean cosine gap to the hardest same-identity opposite-view feature, using the current batch and detached queue. $\mathcal{L}_d$ applies the same alignment to native concatenated descriptors using eligible current-batch pairs, excluding rows assigned a queue donor. For minimum cross-view positive similarity s_i^+ and maximum cross-view negative similarity s_i^-, $\mathcal{L}_n = \mathrm{mean}_i[0.03 + s_i^- - s_i^+]_+$, where $[a]_+ = \max(a, 0)$. The contrastive term uses native descriptors, $s_{ij} = \cos(f_i, \mathrm{sg}(f_j))$, identity labels y_i, view labels ν_i, and opposite-view keys from the current batch, averaging over rows with a cross-view positive:

$$\mathcal{L}_c = \mathrm{mean}_i \left[\log \sum_{j:\nu_j \neq \nu_i} \exp(s_{ij}/0.07) - \log \sum_{j:\nu_j \neq \nu_i,\, y_j = y_i} \exp(s_{ij}/0.07) \right]. \tag{19}$$

Rows lacking a required positive or negative set are omitted, and auxiliary terms are zero when their accepted or eligible set is empty. Keys and queued features are detached. The base objective is

$$\begin{aligned} \mathcal{L}_{\mathrm{base}} = \mathcal{L}_{\mathrm{ID}} + \mathcal{L}_{\mathrm{tri}} + 0.10\mathcal{L}_a + 0.15\mathcal{L}_d + 0.10\mathcal{L}_n + 0.15\mathcal{L}_c \\ + 0.08q\left(\mathcal{L}_{\mathrm{replay,ID}} + \mathcal{L}_{\mathrm{replay,tri}} + 0.25\mathcal{L}_{\mathrm{teacher}}\right). \end{aligned} \tag{20}$$

Replay classification and triplet supervision each use the stream average $(L_{\mathrm{raw}} + L_{\mathrm{proj}} + 0.25L_{\mathrm{spatial}})/3$, with spatial feature $[q_P(S) + c^s]/2$ and selected class token c^s. Teacher consistency averages the three stream-wise cosine gaps to detached clean descriptors, with outer weight 0.25.

The auxiliary objectives are

$$\mathcal{L}_{\mathrm{GVPS}} = \lambda_t \mathcal{L}_{\mathrm{topo}} + \mathcal{L}_{\mathrm{boundary}} + \lambda_c \mathcal{L}_{\mathrm{confidence}} + \lambda_s \mathcal{L}_{\mathrm{support}}. \tag{21}$$

$$\begin{aligned} \mathcal{L}_{\mathrm{OSIC}} &= \mathcal{L}_{\mathrm{metric}} + \lambda_e \mathcal{L}_{\mathrm{evidence}} + \lambda_h \mathcal{L}_{\mathrm{hard}}, \\ \mathcal{L}_{\mathrm{metric}} &= \mathrm{mean}_b \left[1 - \cos(f_{o,b}, f_{s,b})\right], \\ \mathcal{L}_{\mathrm{evidence}} &= \frac{\sum_{b,i} w_{b,i}\left[1 - \cos(o^f_{b,i}, \mathrm{sg}(\widetilde{s}_{b,i}))\right]}{\max(1, \sum_{b,i} w_{b,i})}. \end{aligned} \tag{22}$$

Here, f_o and f_s are final clean and synthetic 1280-dimensional descriptors, b indexes accepted source observations, and sg denotes stop-gradient.

We use $\lambda_t = 0.15$, $\lambda_c = 0.1$, $\lambda_s = 0.1$, $\lambda_e = 0.25$, and $\lambda_h = 0.1$, with outer weights $\lambda_{\mathrm{G}} = 0.05$ and $\lambda_{\mathrm{O}} = 0.10$ in Equation 8.

GVPS topology averages the mean elementwise absolute differences of $S - O$ along horizontal and vertical grid neighbors, normalized by the detached clean root-mean-square feature scale. The boundary term averages raw and projected cosine-gap hinges above 0.04. Confidence averages $[0.70 - p_y]_+$ over the two final identity heads. Support averages $[0.50 - \mathrm{mean}_i\, w_i]_+$ over accepted observations, using straight-through support $\bar{w}_i = w_i + \ell_i - \mathrm{sg}(\ell_i)$, where $\ell_i = [1 + \cos(o_i, \widetilde{s}_i)]/2$, to preserve the forward value while retaining a semantic gradient.

OSIC evidence consistency aggregates accepted token positions with detached support weights. For clean and synthetic distances d_o^f, d_s^f to the detached donor in final descriptor space, the hard-positive term averages over observations with an available donor:

$$\mathcal{L}_{\mathrm{hard}} = \mathrm{mean}\left([d_o^f + 0.004 - d_s^f]_+ + [d_s^f - 0.65]_+\right). \tag{23}$$

The absolute final-descriptor bound 0.65 differs from the relative difficulty-increment bound 0.007 in HPPS selection space.

Table 4: Method-specific defaults. Basic training settings appear in Appendix C.2.

Setting	Value
Weight decay / norm-parameter decay	10^{-4} / 5×10^{-4}
Gradient norm cap / mixed precision	5 / disabled
Training queue capacity	2048 detached feature observations
Candidate count / candidate scales	5 / evenly spaced in $[0.35, 1]$
Nominal patch transfer / vanish fraction	0.10 / at most 0.05
Class-token mixing strength	$0.02q(p)\ell_k$
Distractor row / token fraction	0.15 / 0.10
Distractor residual coefficient	$0.40q(p)\ell_k$
DPEF relative support threshold / residual scale	0.82 / 0.10

Settings. Images undergo horizontal flipping and random erasing, each with probability 0.5, and random cropping after 10-pixel padding. RGB channels are normalized by subtracting and dividing by 127.5. The CARGO loader contains 51,451 training images from 2,500 identities. Raw query/gallery counts are 312/50,712 (ALL), 134/32,268 (A→G), 134/18,444 (A→A), and 178/32,268 (G→G). Evaluation excludes same-identity same-camera gallery matches and skips queries without a valid match.

C Additional Experimental Details

This appendix summarizes the benchmark protocols, training settings, ablation controls, efficiency, additional result analyses, and parameter sensitivity settings of GPS.

C.1 Benchmark Protocols

All four benchmarks use their official training and test splits. AG-ReID v1 (Nguyen et al., 2023) contains 388 identities and supports A→G and G→A retrieval. AG-ReID v2 (Nguyen et al., 2024) contains 100,502 images of 1,615 identities from aerial (A), CCTV (C), and wearable (W) platforms, evaluated under A→C, C→A, A→W, and W→A. CARGO (Zhang et al., 2024b) contains 108,563 synthetic images of 5,000 identities and supports ALL, A→G, A→A, and G→G retrieval. LAGPeR (Wang et al., 2025a) contains 63,841 captured images of 4,231 identities, evaluated under A→G, G→A, and G→AG. Evaluation follows each benchmark's camera exclusions and valid-query rules. Our CARGO A→G evaluation uses aerial queries and a ground gallery.

C.2 Training Settings

We implement GPS in PyTorch with FastReID (He et al., 2023) and train on one NVIDIA H200 NVL GPU. The primary CLIP ViT-B/16 model uses 256×128 inputs and a 1280-dimensional retrieval descriptor. View-balanced batches contain 64 images, four per identity. Augmentation comprises horizontal flipping, ten-pixel padding with random cropping, and random erasing. The method stage uses Adam for 60 epochs, with learning rates of 10^{-5} for general parameters and 5×10^{-6} for pretrained parameters, 100 linear warm-up iterations, and cosine decay to 10^{-6}. The CLIP controls share the encoder and sampler. Method-specific defaults appear in Table 4.

C.3 Ablation Settings

CARGO is the primary ablation benchmark. Plain CLIP uses its native identity and triplet objectives. The module ablations share the clean-image terms defined in Appendix B.4. The CMV study evaluates all eight operator combinations. Disabled operators use identity maps, while active operators follow the order C, M, V and recompute local responses after each edit. Donor selection, HPPS, DPEF, and loss weights remain fixed. Thus, the all-off CMV reference retains donor-based synthesis, whereas the module baseline disables all synthetic branches.

Without HPPS, we use the middle candidate level with the shared identity and validity checks. Without DPEF, we remove its pooled residual while retaining global mixtures and synthetic supervision.

Each module is evaluated together with its dependent objectives. The loss ablation retains the full architecture and varies only the outer weights of GVPS and OSIC, covering neither objective, either one, and both. These controls isolate the contribution of the operators, the use of synthesized evidence, and the auxiliary supervision.

C.4 Efficiency

Table 5 reports trainable parameters and average training time per epoch.

Table 5: Trainable parameters and average training time per epoch. Plain CLIP includes the ReID heads. The symbol + denotes the parameter increment from GPS.

Method	Trainable parameters (M)	Training time (s/epoch)
Plain CLIP (Radford et al., 2021)	89.344	120.81
GPS	+0.590	164.57

GPS adds 0.590M trainable parameters and increases training time from 120.81 to 164.57 s per epoch. The 589,826 parameters comprise a 768×768 DPEF projection and two scalar gates. The additional computation comes from candidate synthesis, selection, fusion, and supervision. All five candidates share one encoder pass per batch, and inference retains the original-image descriptor path.

C.5 Additional Result Analysis

The stronger cross-view gain on CARGO is consistent with the motivation of GCPS. Elevated viewpoints change the spatial distribution, relative contribution, and visibility of local identity cues. Geometry-guided synthesis exposes the encoder to these changes through hard positives constrained by identity consistency, encouraging it to recognize an identity when familiar local evidence is reorganized or weakened. This training objective supports matching under the pronounced appearance changes between aerial and ground observations. The improvements across synthetic and captured imagery further suggest that learning from such altered evidence is useful beyond a particular camera pairing or image source. Within-view gains also indicate preserved identity discrimination.

C.6 Additional Ablation Analysis

Compression, merging, and vanishing. The complementary gains in Table 2(a) reflect the different changes in local evidence induced by aerial–ground viewpoints. Compression and merging expose the model to altered spatial distributions and contributions of identity cues, while vanishing encourages recognition from the remaining evidence when local details become unavailable. Their differing individual effects and consistent joint gains support learning these changes together. The resulting hard positives cover a wider range of local appearances, helping identity learning draw on cues that remain useful as viewpoint changes.

GCPS, HPPS, and DPEF. The pattern in Table 2(b) supports synthesis as the central mechanism, with selection and fusion improving how the generated positives contribute to learning. HPPS controls the challenge presented to the current model within identity-consistency constraints, while DPEF emphasizes evidence supported by both observations. Together, they connect candidate difficulty with the identity evidence used for training.

GVPS and OSIC. GVPS encourages coherent edits with sustained identity support, while OSIC connects the altered observation to its original identity representation. Their complementary benefits support coupling the construction of challenging positives with supervision that keeps those changes useful for identity matching.

C.7 Parameter Sensitivity Settings

Figure 3 evaluates eight settings each for GCPS and HPPS and seven for DPEF. GCPS varies the strength coefficient in $s_k = 0.25q(p)\ell_k$, whose default is 0.25. HPPS varies $\delta_{\min}$ with interval width $\delta_{\max} - \delta_{\min}$ fixed at 0.003. Its default lower bound is 0.004, giving the default upper bound 0.007. DPEF varies the coefficient of the pooled residual in the global update described in Appendix B.3, with default 0.10. Horizontal axes show the tested values. Bars show mAP and Rank-1, and the line shows their average.

D Extended Related Work

Broader observation conditions. Multi-modal multi-platform ReID and WHU-MARS combine platform variation with sensor differences (Ha et al., 2025; Zhao et al., 2026), extending identity matching beyond aerial-ground viewpoint shifts. Shared and observation-specific feature decomposition also supports multimodal object re-identification (Feng et al., 2025). These studies complement view-aware AG-ReID by considering additional sources of appearance variation.

Local representation and correspondence. Cross-view methods organize shared identity information and view-dependent appearance through alignment and decomposition (Huang et al., 2026; Mao et al., 2026). Part-based pooling and global–local aggregation preserve fine-grained identity cues (Sun et al., 2021; Wang et al., 2018), while TransReID models patch interactions with camera and view information (He et al., 2021). Orientation and occlusion make local correspondence dependent on visibility (Sun & Zheng, 2019; Miao et al., 2019; Wang et al., 2020; Huang et al., 2023). Pose-guided and adaptive region selection address this variation (Qiu et al., 2024; Wang & Pishgar, 2025; Zhang et al., 2024a; Wu et al., 2024). SeCap recalibrates visual prompts, and ViSA uses expert-conditioned queries and patch relations to associate viewpoint-dependent evidence (Wang et al., 2025a; Zhang et al., 2026).

Representation and semantic guidance. High-resolution and hierarchical representations retain spatial detail, while token interactions and feature fusion support informative region selection (Wang et al., 2021; Liu et al., 2021; Wang et al., 2024b; 2025b). Semantic guidance draws on information shared across appearances (Li & Gong, 2023; Zhu et al., 2026). CLIP provides image–text pretraining (Radford et al., 2021), which CLIP-ReID adapts through identity-specific text tokens. Attribute-conditioned approaches add person attributes and view information (Li et al., 2023; Zhang et al., 2025). Description-based guidance extends these semantics to visible-infrared ReID (Hu et al., 2024). In text-to-aerial person retrieval, fuzzy token alignment weights uncertain associations under missing or noisy visual cues (Deng et al., 2026). These complementary sources of guidance support semantic matching across heterogeneous observations.

Task-specific augmentation. Task-specific transformations connect augmentation to appearance variation (Chen et al., 2022; Wang et al., 2024a). RLE applies local linear transformations in cross-spectral ReID (Tan et al., 2024). For AG-ReID, VIF-AGReID's Patch-Level RotateMix rotates and blends image patches with spatially varying probabilities, together with an angular learning objective (Khalid et al., 2025). Its transformation policy is not conditioned on the relative difficulty of an observed same-identity cross-view pair.

Conditional person synthesis. Synthetic training data diversifies the observations available for identity learning (Zhang et al., 2021b). Pose-normalized generation creates person images under prescribed poses (Qian et al., 2018), while 3D-LENS lifts ground-view content into a 3D representation to synthesize elevated views for single-view training (Grolleau et al., 2026). For AG-ReID, SD-ReID combines identity and view conditions with a diffusion prior to enhance person representations (Wang et al., 2026). GPS instead constructs local variations in patch-feature space, using observed same-identity cross-view information to regulate candidate difficulty and evidence fusion.

Hard-positive learning. Hard-positive mining selects challenging same-identity associations (Hermans et al., 2017; Yu et al., 2018), while synthesis can additionally vary the evidence available within them (Huang et al., 2018). GPS therefore combines structured patch-feature synthesis

with identity-constrained selection, choosing the hardest consistent candidate within a prescribed range of additional difficulty relative to the observed pair.

E CROSS-DATASET GENERALIZATION

We train on LAGPeR and evaluate on AG-ReID v1 and AG-ReID v2 under a unified protocol. GPS achieves the highest mAP and Rank-1 in all six directions compared with HiHR, ViSA, and SeCap (Table 6). This consistency supports the transferability of geometry-guided hard-positive learning, in line with GCPS varying local cue organization and visibility to improve robustness under new capture conditions.

Table 6: Cross-dataset generalization from LAGPeR to AG-ReID v1/v2 (mAP and Rank-1, %). Best and second-best results are bold and underlined.

Method	AG-ReID v1				AG-ReID v2							
	A→G		G→A		A→C		C→A		A→W		W→A	
	mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1
SeCap (Wang et al., 2025a)	36.66	53.69	33.09	42.72	27.23	40.45	22.00	28.88	23.51	37.66	23.31	30.90
ViSA (Zhang et al., 2026)	43.42	60.97	41.54	52.29	35.57	50.42	31.18	41.08	30.63	47.80	29.60	39.44
HiHR (Yang & Zhang, 2026)	48.05	63.31	48.20	59.15	42.86	57.51	41.49	52.46	40.08	59.35	37.69	48.46
GPS	**50.56**	**65.08**	**48.95**	**61.54**	**45.09**	**59.55**	**42.16**	**54.17**	**42.32**	**59.48**	**40.75**	**51.97**