This article has been accepted for publication in IEEE Transactions on Information Forensics and Security. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2026.3744610

Rectifying Geometry-Induced Similarity Distortions for *Real-World* Aerial–Ground Person Re-Identification

Kailash A. Hambarde, Hugo Proença, *Senior Member, IEEE*

***Abstract*—Aerial ground person re-identification (AG-ReID) is fundamentally challenged by extreme viewpoint and distance discrepancies between aerial and ground cameras, which induce severe geometric distortions and invalidate the assumption of a shared similarity space across views. The existing methods primarily use geometry aware feature learning or appearance-conditioned prompting, and implicitly assume that the geometry invariant dot-product similarity used within "attention" remains reliable under large viewpoint and distance variations. The main hypothesis in this work is that this assumption does not hold: extreme camera geometry and scale variations systematically distort the query–key similarity space, and heavily compromise attention-based matching, even when feature representations are partially aligned. To address this issue, we introduce the notion of *Geometry-Induced Query–Key Transformation* (GIQT), a lightweight and low-rank module that explicitly rectifies the similarity space by conditioning query-key interactions on camera geometry. Rather than modifying the feature content or the attention formulation itself, GIQT adapts the similarity computation to compensate for dominant, geometry-induced anisotropic distortions. Building on this local similarity rectification, we consider a geometry-conditioned prompt generation mechanism as an auxiliary component that provides global and view-adaptive representation priors, that are derived directly from camera geometry. Our experiments were carried out on four aerial-ground person re-identification benchmarks, where the proposed framework attains state-of-the-art Rank-1 accuracy, together with the best or second-best mAP, and improves robustness under extreme, corrupted, and previously unseen geometric configurations. In practice, this work gives a step towards operation in high-altitude and large viewing-angle regimes. Also, such contribution is provided at minimal computational overhead with respect to competing methods. Code: https://github.com/kailashhambarde/GeoReID.git**

***Index Terms*—Person re-identification, aerial-ground matching, camera geometry, attention, cross-view retrieval, UAV surveillance.**

I. Introduction

PERSON re-identification (ReID) is a classical problem of computer vision learning domains, and aims to match identities across non-overlapping cameras [1]–[3]. Here, most significant progresses have been achieved in ground-to-ground ReID, using emerging network architectures and training strategies [6]–[8]. However, as ground based camera networks remain limited in spatial coverage, efforts have been concentrated in using unmanned aerial vehicles (UAVs) to broaden surveillance capabilities [10], [11] of such type of systems. The combination of aerial and ground cameras gave rise to the aerial-ground person re-identification (AG-ReID) problem, which requires to match identities across drastically different viewpoints and distances. AG-ReID is substantially more challenging than conventional ReID due to extreme perspective changes: aerial images often exhibit top-down or oblique views, while ground images typically capture frontal or profile appearances. Such viewpoint discrepancies induce severe geometric distortions, including scale compression, foreshortening, and body part displacement, yielding large appearance variation and unreliable spatial correspondence across views [20], [21]. Beyond appearance variation, extreme camera geometry exposes a fundamental limitation of the existing ReID pipelines: the assumption of a shared similarity space across views. Modern transformer based ReID models rely on dot-product attention to compute similarity between local features [15], [51]. Even when feature representations are semantically aligned, viewpoint geometry warps spatial correspondence and feature orientation in a highly anisotropic manner [14], [27]. As a result, the geometry invariant dot-product similarity used within attention becomes unreliable: corresponding regions may yield low similarity, while unrelated regions may spuriously align. Importantly, this does not indicate a limitation of attention or cross attention mechanisms themselves, but rather a mismatch between extreme camera geometry and the default similarity metric used to compute attention weights. This behavior has been implicitly observed in prior AG-ReID studies, where performance degrades sharply as geometric disparity increases [25], [26]. This effect is empirically illustrated in Fig. 1, where performance degrada-

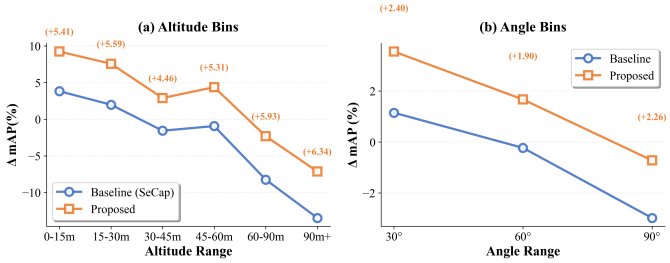


Fig. 1. Comparison between the performance degradation of the well known SeCap method as baseline and our proposal, across altitude and angle bins. The y-axis provide the Δ mAP = (Bin mAP - Overall SeCap A→G mAP). Values above zero indicate *better-than-overall* performance and values below zero indicate performance degradation. It is evident that the proposed method keeps superior performance in extreme geometric regimes.

Manuscript received February XX, 2025; revised XX XX, 2025.

This work was supported by national funds through FCT – Fundação para a Ciência e a Tecnologia, I.P., and, when applicable, co-funded by EU funds under project UID/50008/2025 – Instituto de Telecomunicações.

Kailash A. Hambarde and Hugo Proença are with the Instituto de Telecomunicações and the Department of Computer Science, Universidade da Beira Interior, Portugal.

Corresponding author: Kailash A. Hambarde (e-mail: kailash.hambarde@ubi.pt).

Authorized licensed use limited to: b-on: UNIVERSIDADE DA BEIRA INTERIOR. Downloaded on October 21,2026 at 06:32:31 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Information Forensics and Security. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2026.3744610

tion correlates monotonically with altitude and viewing angle differences. The increasing performance gap under extreme geometry indicates that camera geometry introduces structured distortion in similarity computation rather than random noise.

The existing AG-ReID methods mitigate cross view discrepancies using attribute supervision [25], view disentangled representations [27], or prompt based feature adaptation. While effective in moderate settings, these approaches primarily rely on implicit learning of geometric relationships and typically ignore explicit camera geometry during similarity computation, limiting robustness under extreme or unseen viewpoints.

Motivated by these observations, we argue that robust AG-ReID requires explicitly accounting for geometry induced distortion in cross view similarity computation, rather than further modifying feature representations. To this end, we propose a geometry conditioned similarity alignment framework that incorporates camera geometry into both global representation adaptation and local similarity computation.

The main contributions of this work are four fold:

- We identify similarity space distortion as a dominant failure mode in aerial-ground person re-identification under extreme camera geometry, showing that the geometry invariant similarity assumption underlying attention becomes unreliable even when features are reasonably aligned.
- We propose a geometry conditioned similarity alignment framework that explicitly adapts cross view similarity computation rather than relying solely on feature adaptation.
- We introduce the Geometry-Induced Query-Key Transformation (GIQT), a lightweight and model agnostic module that reshapes the attention similarity space via a geometry conditioned low rank transformation.
- We provide empirical evidence of stronger robustness and generalization under extreme, corrupted, and unseen geometric conditions across multiple AG-ReID benchmarks, attaining state-of-the-art Rank-1 accuracy together with the best or second best mAP. We further analyze the protocol level cases in which the gain is not uniform across both metrics.

II. Related Work

A. Aerial–Ground Person Re-Identification

Aerial-ground ReID extends conventional ReID to scenarios involving UAV mounted and ground based cameras, introducing extreme viewpoint, scale, and orientation discrepancies that fundamentally violate assumptions made in ground-to-ground settings [25], [33]. Unlike conventional ReID, where viewpoint variations are relatively limited, AG-ReID must cope with severe geometric distortions caused by altitude changes, oblique viewing angles, and drastic perspective shifts. Early research in AG-ReID primarily focused on benchmark construction to enable systematic evaluation. Nguyen et al. introduced the AG-ReID.v1 dataset [25] and its extension AG-ReID.v2 [26], highlighting the difficulty of matching identities across aerial and ground views. Zhang et al. proposed the large scale synthetic CARGO benchmark to explicitly stress test cross view appearance changes under extreme geometry [27]. Subsequent video based benchmarks, such as G2A-VReID and AG-VPReID, incorporated temporal cues while retaining the same fundamental geometric challenges [34], [35]. These datasets demonstrate that appearance driven representations are inherently fragile under extreme viewpoint geometry.

B. Implicit Geometry Modelling in AG-ReID

Most existing AG-ReID methods address viewpoint discrepancies through implicit feature learning. Multi stream architectures separate aerial and ground representations and rely on auxiliary supervision, such as identity attributes, to encourage view invariant encoding [26]. Transformer based approaches, including the View Decoupled Transformer (VDT), attempt to disentangle identity related and viewpoint related factors using view tokens and orthogonality constraints [27]. Prompt based methods, such as SeCap, further adapt representations through appearance conditioned prompts or dynamic token selection [31]. While effective in practice, these approaches learn geometric relationships implicitly from data. Viewpoint geometry such as camera altitude, viewing angle, or sensor configuration, is not explicitly modelled, limiting interpretability and control over how geometric distortions influence feature alignment. As a result, such methods often struggle to generalize under unseen or extreme viewpoint configurations.

C. Geometry Aware and View Invariant Learning

Several studies have explored mechanisms related to geometric robustness in cross view recognition. Adversarial domain adaptation has been employed to reduce distribution gaps between aerial and ground features [36], while view aware attention mechanisms attempt to capture view dependent transformations implicitly [27]. Diffusion based synthesis augments training data with generated cross view samples [37], and disentangled representation learning separates identity from view specific factors [26]. Rotation robustness has also been investigated through equivariant architectures and rotation aware tokens [38]–[40]. However, these approaches primarily rely on data augmentation or implicit representation learning and do not explicitly incorporate known geometric priors. Crucially, readily available metadata in aerial sensing, such as altitude, viewing angle, and camera identity, are typically ignored during both representation learning and attention based similarity computation.

D. Geometric Alignment and Multi-Modal Cues

Geometric and semantic alignment methods aim to reduce cross view mismatch by establishing correspondences across views. GSAlign leverages coupled geometric semantic alignment networks to learn dense correspondences [41], while part based and keypoint driven approaches rely on pose estimation to guide alignment [42], [43]. However, keypoint detection and part localization are unreliable in low resolution aerial imagery and under severe occlusion, limiting their effectiveness in AG-ReID scenarios. Multi modal approaches introduce textual or language based supervision to enhance

Authorized licensed use limited to: b-on: UNIVERSIDADE DA BEIRA INTERIOR. Downloaded on October 21,2026 at 06:32:31 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Information Forensics and Security. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2026.3744610

view invariance [44]–[46]. While promising, such methods increase system complexity and often depend on additional annotations or pretrained language models, making them less suitable for lightweight UAV based deployment. Despite significant progress, existing AG-ReID approaches share several limitations. First, geometric relationships are typically modeled implicitly, limiting interpretability and controllability. Second, robustness degrades substantially under extreme altitude and viewing angle variations. Third, many methods introduce considerable computational overhead, hindering real time UAV deployment. Finally, generalization across datasets with different camera geometries remains limited. These limitations motivate the need for explicitly geometry aware AG-ReID frameworks that directly incorporate aerial sensing priors into both global representation adaptation and local similarity computation, rather than relying solely on implicit feature learning.

III. Method

A. Overview

The overall framework of the proposed geometry conditioned similarity alignment framework, as illustrated in Fig. 2 (a), adopts an encoder-decoder transformer architecture. The encoder is the View Decoupling Transformer (VDT) [27]. In contrast to the conventional ViT [32], our approach incorporates the **View** token and performs hierarchical decoupling of the **Cls** token at each layer, effectively segregating view related and view invariant features within the **Cls** token, while extracting local features from the input [31]. The decoder comprises the Geometry Conditioned Prompt Generation module (**GCPG**) and the Cross View Feature Transformation module (**CVFT**). The **GCPG** adaptively generates prompts for different viewpoints based on the view invariant descriptor X_{inv} and the camera geometry embedding $\mathbf{e}_{\text{geo}}$. The **CVFT** utilizes the geometry conditioned prompts P_{geo} and applies the Geometry Induced Query Key Transformation (**GIQT**) inside cross attention to decode geometry aligned local features. The final ReID representation is $\mathbf{Out} = [X_{\text{inv}}, X_{\text{ref}}]$, i.e., view invariant global descriptor plus refined local features; the encoder outputs $\mathbf{Cls}$, $\mathbf{View}$, and X_{local} are used only internally as in Eq. (1). The overall framework can be described as follows:

$$\begin{aligned}
[\mathbf{Cls}, \mathbf{View}, X_{\text{local}}] &= \text{VDT}([\mathbf{CLS}, \mathbf{View}, \text{tokenization}(\mathbf{X})]),\\
X_{\text{inv}} &= \mathbf{Cls} - \mathbf{View},\\
\mathbf{e}_{\text{geo}} &= [\mathbf{e}_{\text{cam}};\, \mathbf{e}_{\text{alt}};\, \mathbf{e}_{\text{angle}}],\\
P_{\text{geo}} &= \mathcal{G}_{\text{prompt}}(X_{\text{inv}}, \mathbf{e}_{\text{geo}}),\\
X_{\text{ref}} &= \mathcal{G}_{\text{align}}(X_{\text{local}}, P_{\text{geo}}, \mathbf{e}_{\text{geo}}),\\
\mathbf{Out} &= [X_{\text{inv}}, X_{\text{ref}}],
\end{aligned} \tag{1}$$

where VDT is the View Decoupling Transformer; $\mathbf{Cls}$ and $\mathbf{View}$ are the class token and view token output by the encoder; $\text{tokenization}(\mathbf{X})$ denotes the process of converting the input image $\mathbf{X}$ into patch tokens; $X_{\text{local}} \in \mathbb{R}^{B\times N\times d}$ denotes the local patch features (B batch size, N number of patches, d feature dimension); X_{inv} is the view invariant global descriptor; $\mathbf{e}_{\text{geo}}$ is the geometry embedding encoding altitude, viewing angle, and camera identity; P_{geo} denotes the geometry conditioned prompts; $\mathcal{G}_{\text{prompt}}$ is the prompt generation module (GCPG); $\mathcal{G}_{\text{align}}$ is the local refinement module (CVFT) that uses GIQT; X_{ref} is the refined local features; and $\mathbf{Out}$ is the final representation used for ReID.

B. Geometry Metadata Acquisition

Our framework relies on camera geometry cues: altitude, viewing angle, and camera identity. When such metadata is provided by the dataset, we directly encode the ground truth geometry. When metadata is unavailable or incomplete, we train a vision only multi task geometry prediction network to estimate these cues from RGB images. As shown in Fig. 3, a shared ResNet-50 encoder extracts visual features, followed by task specific heads for altitude and distance regression and viewing angle classification. The predicted geometry is used consistently during training and inference, enabling geometry conditioned alignment without external sensors.

C. Geometry Conditioned Prompt Generation (GCPG)

Global cross view alignment requires adapting high level representation priors to camera geometry. To this end, we introduce a geometry conditioned prompt generation module that maps the view invariant descriptor and geometry embedding to global prompt offsets. As illustrated in Fig. 2 (b), the module first encodes geometry cues into a single embedding vector. Geometry cues are encoded as:

$$\mathbf{e}_{\text{geo}} = [\mathbf{e}_{\text{cam}};\, \mathbf{e}_{\text{alt}};\, \mathbf{e}_{\text{angle}}] \in \mathbb{R}^{d_{\text{geo}}}, \tag{2}$$

where $\mathbf{e}_{\text{cam}} \in \mathbb{R}^{d_{\text{cam}}}$ is the camera identity embedding from a learnable table indexed by camera ID c; $\mathbf{e}_{\text{alt}} \in \mathbb{R}^{d_{\text{alt}}}$ is the altitude embedding obtained by discretizing altitude h into N_{alt} bins and indexing a learnable table $\mathbf{E}_{\text{alt}}$; $\mathbf{e}_{\text{angle}} \in \mathbb{R}^{d_{\text{angle}}}$ is the viewing angle embedding from discretizing angle θ into N_{angle} bins and indexing $\mathbf{E}_{\text{angle}}$; and $d_{\text{geo}} = d_{\text{cam}} + d_{\text{alt}} + d_{\text{angle}}$. Given the view invariant descriptor X_{inv} and geometry embedding $\mathbf{e}_{\text{geo}}$, geometry conditioned prompts are generated as:

$$P_{\text{geo}} = P_{\text{base}} + \alpha \cdot f_{\text{geo}}(X_{\text{inv}}, \mathbf{e}_{\text{geo}}), \tag{3}$$

where $P_{\text{base}} \in \mathbb{R}^{L\times d}$ denotes learnable base prompts (L prompt length, d feature dimension); α is a learnable scaling factor; and $f_{\text{geo}}(\cdot)$ is a lightweight geometry conditioned mapping network. The residual formulation preserves identity discriminative semantics while injecting geometry as a structured bias. Geometry conditioned prompts act as global priors that guide the decoder toward geometry consistent cues under varying viewpoints.

D. Geometry Induced Query Key Transformation (GIQT)

While geometry conditioned prompts provide coarse global adaptation, extreme viewpoint changes introduce local geometric distortions that challenge the geometry invariant similarity assumption used in standard attention. The proposed Geometry Induced Query Key Transformation addresses this by adapting the similarity metric used to compute attention weights. Empirical analysis of cross view feature covariance reveals

Authorized licensed use limited to: b-on: UNIVERSIDADE DA BEIRA INTERIOR. Downloaded on October 21,2026 at 06:32:31 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Information Forensics and Security. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2026.3744610

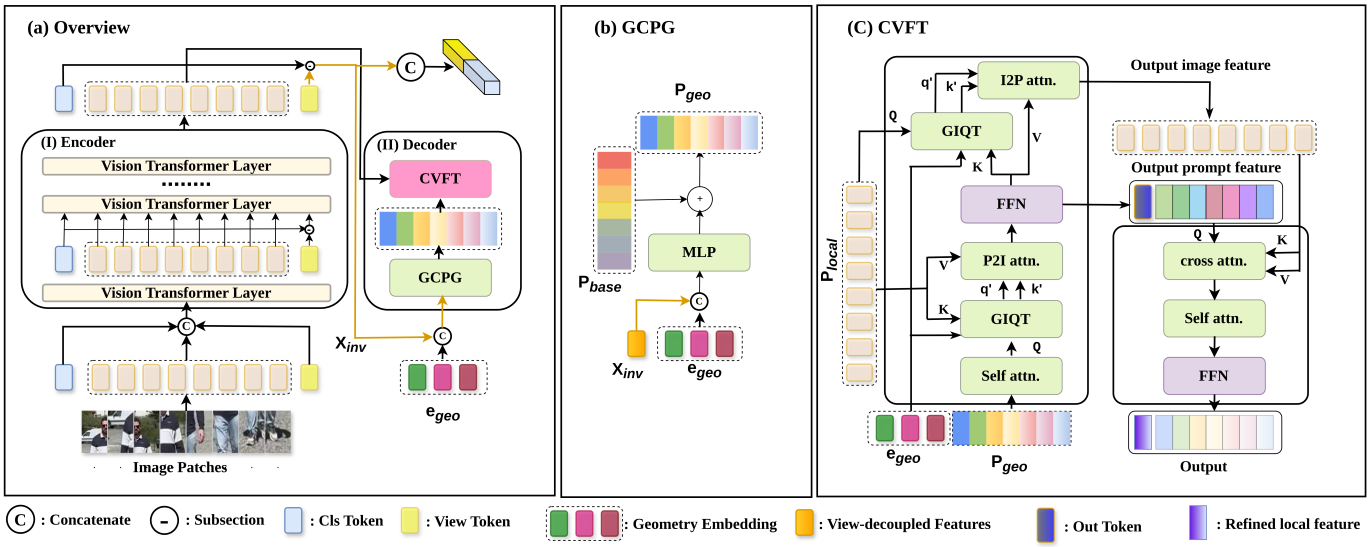


Fig. 2. (a) The architecture of the proposed geometry conditioned similarity alignment framework. The encoder (VDT) extracts visual features and decouples view-related and view-invariant components; the decoder comprises GCPG and CVFT. The final ReID representation is $\mathbf{Out} = [X_{\text{inv}}, X_{\text{ref}}]$ (view-invariant global descriptor plus refined local features from CVFT). (b) Geometry Conditioned Prompt Generation (GCPG): inputs are the view-invariant descriptor X_{inv} and geometry embedding $\mathbf{e}_{\text{geo}}$; an MLP f_{geo} yields geometry-conditioned prompts $P_{\text{geo}} = P_{\text{base}} + \alpha \cdot f_{\text{geo}}(X_{\text{inv}}, \mathbf{e}_{\text{geo}})$. (c) Cross-View Feature Transformation (CVFT, LFRM): two-way attention with GIQT decodes discriminative local features from X_{local} using P_{geo} and $\mathbf{e}_{\text{geo}}$.

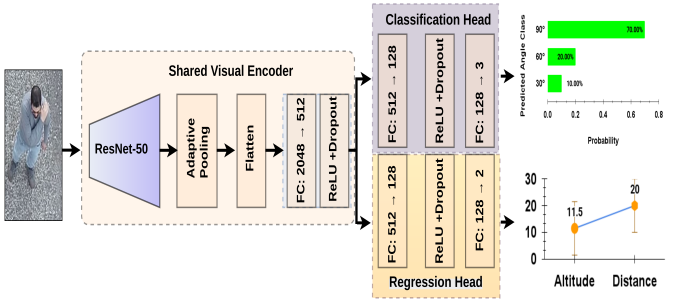


Fig. 3. Vision only multi task geometry prediction. A shared ResNet-50 encoder extracts visual features from RGB images, followed by task specific heads for altitude and distance regression and viewing angle classification.

pronounced spectral decay, indicating that viewpoint induced distortions are highly anisotropic; a small number of dominant components capture most geometry induced variation, motivating a low rank correction strategy. Before computing attention, geometry conditioned transformations are applied to queries and keys and values are kept unchanged to preserve feature content:

$$Q' = T_Q(\mathbf{e}_{\text{geo}})Q, \qquad K' = T_K(\mathbf{e}_{\text{geo}})K. \tag{4}$$

Where $Q \in \mathbb{R}^{L\times d_k}$ and $K \in \mathbb{R}^{N\times d_k}$ are the query and key matrices per head ($d_k = d/H$ is the head dimension, H the number of heads); $T_Q(\mathbf{e}_{\text{geo}})$ and $T_K(\mathbf{e}_{\text{geo}}) \in \mathbb{R}^{d_k\times d_k}$ are learned transformation matrices conditioned on $\mathbf{e}_{\text{geo}}$; and Q', K' are the transformed query and key. In implementation, T_Q and T_K are applied *per head* to the head dimension vectors. Attention is then computed as $\text{Attention}(Q', K', V) = \text{softmax}(Q'(K')^\top/\sqrt{d_k})V$, with V unchanged. To avoid over parameterization and stabilize training, we adopt a low rank residual formulation:

$$T(\mathbf{e}_{\text{geo}}) = I + U(\mathbf{e}_{\text{geo}})V(\mathbf{e}_{\text{geo}})^\top. \tag{5}$$

Where I is the identity matrix; U and V are predicted from $\mathbf{e}_{\text{geo}}$ via a lightweight geometry encoder and per head linear predictors, with $U, V \in \mathbb{R}^{d_k\times r}$ and rank $r \ll d_k$. This formulation targets dominant geometry sensitive directions. The transformation can be applied efficiently without materializing $d_k \times d_k$ matrices: for a vector $\mathbf{q}$, $\mathbf{q}' = \mathbf{q} + U(V^\top\mathbf{q})$. Optionally, per head residual gating blends transformed and original features: $\mathbf{q}_{\text{out}} = \alpha^{(h)}\mathbf{q}' + (1-\alpha^{(h)})\mathbf{q}$, where $\alpha^{(h)} = \sigma(\beta^{(h)})$ is a learnable scalar per head. By adapting the comparison space rather than feature content, GIQT suppresses unstable viewpoint dependent similarity directions and emphasizes geometry consistent cues.

E. Geometry Aligned Local Feature Refinement (CVFT)

The Cross View Feature Transformation (CVFT), is the transformer based decoder, as illustrated in Fig. 2 (c). It extracts geometry aligned discriminative features from the local features X_{local} using the geometry conditioned prompts P_{geo} from the GCPG and the GIQT inside cross attention. The CVFT consists of the two way attention module and the feature fusion module. The two way attention module employs self attention (SA) and cross attention (CA) in both prompt-to-image and image-to-prompt directions; all cross attention operations use GIQT transformed queries and keys. The two way attention module can be described as follows:

$$\begin{aligned} F_{\text{P}} &= \text{FFN}(\text{CA}(\text{SA}(P_{\text{geo}}), X_{\text{local}}, X_{\text{local}})) + P_{\text{geo}}, \\ F_{\text{I}} &= \text{CA}(X_{\text{local}}, F_{\text{P}}, F_{\text{P}}) + X_{\text{local}}, \end{aligned} \tag{6}$$

Authorized licensed use limited to: b-on: UNIVERSIDADE DA BEIRA INTERIOR. Downloaded on October 21,2026 at 06:32:31 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Information Forensics and Security. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2026.3744610

where F_{P} represents the prompt features output by the two way attention module; F_{I} denotes the image (local) features output by the two way attention module; FFN stands for the Feed Forward Network; CA is the cross attention mechanism (with GIQT applied to queries and keys before the attention score computation); SA is the self attention mechanism; X_{local} signifies the local patch features from the encoder; and P_{geo} refers to the geometry conditioned prompts. The feature fusion module employs cross attention and self attention to integrate the output token with the prompt and image features, decoding the final local representation. The feature fusion module can be described as follows:

$$[\mathbf{Out}, \_] = \mathrm{FFN}(\mathrm{SA}(\mathrm{CA}([\mathbf{Out}_{\mathrm{token}}, F_{\mathrm{P}}], F_{\mathrm{I}}, F_{\mathrm{I}}))), \quad (7)$$

where $\mathbf{Out}$ is the final local descriptor used for ReID; $\mathbf{Out}_{\mathrm{token}}$ is a learnable output token; F_{P} is the prompt features; and F_{I} is the refined image features. The resulting refined features are explicitly aligned with respect to camera geometry, yielding robust local representations under extreme cross view variations.

Relation to SeCap LFRM. The two way attention and feature fusion structure of CVFT follows SeCap's Local Feature Refinement Module (LFRM) [31]. The distinction is confined to the similarity computation: when GIQT is removed, every geometry conditioned transform reduces to the identity ($T(\mathbf{e}_{\mathrm{geo}})=I$, i.e. $UV^{\top}=0$), the queries and keys are left unchanged, and CVFT becomes *exactly* the SeCap LFRM. GeoReID therefore does not restructure the decoder; it modifies only *how* tokens are compared inside cross attention. The component ablations (Table VI) quantify this residual: the "+GIQT" rows isolate the gain attributable solely to geometry conditioned similarity rectification over the SeCap equivalent decoder.

F. Optimization

In addition to the ID classification loss and the triplet loss commonly used in ReID tasks, our loss functions are further enhanced with the view classification loss, the orthogonality loss, and the geometry conditioned prompt regularization [31].

View Classification Loss: To achieve decoupling of view related features, we utilize a view classifier and employ the view classification loss for constraint. The loss function can be formalized as follows:

$$\mathcal{L}_{\mathrm{view}} = -\sum_{i=1}^{N} y_i \log(p_i), \quad (8)$$

where y_i is the ground truth view label for the i-th sample; p_i is the predicted probability of the i-th sample belonging to the correct view class; and N is the total number of samples in the batch.

Orthogonality Loss: To ensure thorough view decoupling, we introduce the orthogonality loss. The expression is as follows:

$$\mathcal{L}_{\mathrm{orth}} = \sum_{i=1}^{d} \big|\langle \mathbf{inv}_i, \mathbf{v}_i \rangle\big|, \quad (9)$$

where $|\langle \cdot, \cdot \rangle|$ denotes the absolute value of the dot product; d represents the feature dimensionality; $\mathbf{inv}_i$ and $\mathbf{v}_i$ denote the i-th component of the view invariant and view related feature vectors, and the sum runs over the feature dimension d.

Geometry Conditioned Prompt Regularization: To stabilize geometry conditioned prompt adaptation, an ℓ_2 regularization term is applied to the prompt offset:

$$\mathcal{L}_{\mathrm{geo}} = \big\| f_{\mathrm{geo}}(X_{\mathrm{inv}}, \mathbf{e}_{\mathrm{geo}}) \big\|_2^2, \quad (10)$$

where f_{geo} is the geometry conditioned mapping network and the offset is the generated prompt component before adding to P_{base}.

Overall Loss: We apply the ID classification loss and the Triplet loss to both global and local features. The overall optimization objective can be summarized as follows:

$$\begin{aligned} \mathcal{L} = \alpha\big(\mathcal{L}_{\mathrm{ID}}^{\mathrm{global}} + \mathcal{L}_{\mathrm{Tri}}^{\mathrm{global}}\big) + \beta\big(\mathcal{L}_{\mathrm{ID}}^{\mathrm{local}} + \mathcal{L}_{\mathrm{Tri}}^{\mathrm{local}}\big) \\ + \lambda\big(\mathcal{L}_{\mathrm{view}} + \mathcal{L}_{\mathrm{orth}}\big) + \gamma \mathcal{L}_{\mathrm{geo}}, \end{aligned} \quad (11)$$

where α and β are hyperparameters that balance the optimization objectives of global and local features; λ balances view disentanglement (view classification and orthogonality) with identity learning; and γ controls the geometry conditioned prompt regularization. We set $(\alpha, \beta, \lambda, \gamma) = (1.0, 1.0, 0.5, 0.1)$ in all experiments.

IV. Experiments

A. Datasets and Evaluation Protocols

Datasets. We evaluate our approach on four aerial-ground person ReID benchmarks: AG-ReIDv1 [25], AG-ReIDv2 [26], CARGO [27], and DetReIDX [28]. AG-ReIDv1 and AG-ReIDv2 provide ground truth camera geometry annotations, including drone altitude and camera identity. DetReIDX provides drone altitude, viewing angles, and camera identity. In contrast, CARGO only provides camera identity and does not include explicit geometric metadata such as altitude or viewing angle. The CARGO dataset contains 108,563 images of 5,000 identities captured by eight ground cameras and five aerial cameras. AG-ReIDv1 consists of 21,983 images of 388 identities collected using one aerial camera and one ground camera, with aerial views captured at altitudes ranging from 15 to 45 meters. AG-ReIDv2 extends AG-ReIDv1 by introducing additional viewpoints, more identities, and more diverse camera configurations. DetReIDX contains 553 identities and includes one training split and three evaluation protocols: A→A (105,478 images), A→G (72,033 images), and G→A (72,033 images). The dataset is collected across six international university campuses and features substantial variation in drone altitude and viewing angle. For CARGO, we use the proposed vision only geometry predictor to generate altitude, horizontal distance, and viewing angle. For AG-ReIDv1 and AG-ReIDv2, the geometry predictor is used only to estimate viewing angles when ground truth angle annotations are unavailable.

Evaluation Protocols. Following common practice in aerial-ground person ReIDs we evaluate all methods using the Cumulative Matching Characteristic (CMC) at Rank-1 [64] and mean Average Precision (mAP) [65]. For the CARGO dataset, we follow the benchmark defined protocols and report results under four settings: *ALL*, which evaluates performance

Authorized licensed use limited to: b-on: UNIVERSIDADE DA BEIRA INTERIOR. Downloaded on October 21,2026 at 06:32:31 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Information Forensics and Security. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2026.3744610

across all query–gallery pairs; $G\leftrightarrow G$, which evaluates ground-to-ground matching; $A\leftrightarrow A$, which evaluates aerial-to-aerial matching; and $A\leftrightarrow G$, which evaluates the most challenging aerial-to-ground matching scenario. For AG-ReIDv1, we adopt the standard bidirectional cross view evaluation protocol between aerial and ground cameras, namely $A\rightarrow G$ and $G\rightarrow A$. AG-ReIDv2 extends AG-ReIDv1 by introducing additional camera viewpoints. In particular, the wearable view (W) represents a low viewpoint close to human eye level, while the ground view (G) corresponds to fixed CCTV-style cameras. Based on these views, AG-ReIDv2 defines four cross view protocols: $A\rightarrow G$, $G\rightarrow A$, $A\rightarrow W$, and $W\rightarrow A$. For DetReIDX, we follow the official evaluation protocol and report results under three settings: $A\rightarrow A$, $A\rightarrow G$, and $G\rightarrow A$. These protocols involve significant variation in drone altitude, viewing angle, and image resolution, making cross view retrieval particularly challenging. In all cases, A denotes aerial views, G denotes ground views, and W denotes wearable views. The arrow indicates the retrieval direction from query to gallery, while the double arrow indicates bidirectional evaluation. All methods are evaluated using identical train test splits and metrics to ensure fair and reproducible comparison.

B. Implementation Details

Our proposed method is implemented using PyTorch and the FastReID framework. All experiments are conducted on NVIDIA A40 GPUs. We adopt a pre-trained Vision Transformer Base (ViT-B/16) as the backbone network, initialized with ImageNet pre-trained weights. All input images are resized to $256 \times 128 \times 3$ pixels. For data augmentation, we apply random horizontal flipping, padding, and random erasing [67] with a probability of 0.5. The model is trained using a mini batch size of 128 (or 192 for larger datasets), where each batch consists of 32 identities with 4 instances per identity (or 8 instances for datasets with more samples per identity). The model is optimized using SGD [66] with a base learning rate of 0.008 for 120 epochs on the dataset size. We employ a cosine annealing learning rate schedule with a minimum learning rate of 1.6×10^{-6} and a warmup period of 2000 iterations. The weight decay is set to 0.0001, and gradient clipping is enabled for training stability. Mixed precision training (Automatic Mixed Precision) is utilized to accelerate training and reduce memory consumption.

C. Comparison with State-of-the-art Methods

We evaluate the proposed GeoReID framework on four aerial-ground person ReID benchmarks, namely AG-ReIDv1, AG-ReIDv2, CARGO, and DetReIDX, and compare it with a wide range of state-of-the-art methods. Quantitative results on AG-ReID are reported in Table I. GeoReID achieves the highest Rank-1 accuracy under both A↔G and G↔A evaluation protocols. Under the A↔G setting, GeoReID attains 87.02% Rank-1 accuracy and 79.46% mAP, improving upon the strongest prior Rank-1 result while maintaining competitive retrieval quality. Under the G↔A setting, GeoReID further achieves 90.64% Rank-1 accuracy and 83.55% mAP, surpassing previous methods in Rank-1 performance. These results indicate that geometry conditioned similarity alignment consistently improves cross view matching accuracy. Results on AG-ReIDv2 are summarized in Table II, which reports performance across four cross view protocols (A→G, G→A, A→W, and W→A). GeoReID achieves the best mAP under all four protocols and the best Rank-1 under three of them.

TABLE I
PERFORMANCE COMPARISON ON AG-REID UNDER TWO CROSS VIEW PROTOCOLS. ALL METHODS ARE EVALUATED UNDER IDENTICAL TRAINING AND TESTING SETTINGS. METHODS NOT ORIGINALLY DESIGNED FOR AG-REID ARE ADAPTED FOLLOWING THEIR OFFICIAL IMPLEMENTATIONS. GEOMETRY INFORMATION IS PROVIDED ONLY TO METHODS THAT EXPLICITLY SUPPORT GEOMETRY CONDITIONING. BEST RESULTS ARE HIGHLIGHTED AS **BOLD**, AND SECOND BEST RESULTS ARE HIGHLIGHTED AS UNDERLINED.

Method	Protocol 1: A↔G		Protocol 2: G↔A	
	Rank1	mAP	Rank1	mAP
SBS [47]	73.54	59.77	73.70	62.27
BoT [48]	70.01	55.47	71.20	58.83
OsNet [49]	72.59	58.32	74.22	60.99
VV [50]	77.22	67.23	79.73	69.83
Explain [55]	81.28	72.38	82.64	73.35
ViT [51]	81.47	72.61	82.85	73.39
TrasReID [15]	81.80	73.10	83.40	74.60
PFD [59]	82.30	73.60	82.50	73.90
PHA [18]	79.30	71.30	81.10	72.10
FusionReID [56]	80.40	71.40	82.40	74.20
PCL-CLIP [57]	82.16	73.11	86.90	76.28
DTST [43]	83.48	74.51	84.72	76.05
VDT [27]	82.91	74.44	86.59	78.57
CLIP-ReID [19]	72.61	62.09	74.12	64.19
AG-ReID [25]	82.91	72.38	82.85	73.35
SeCap [31]	84.03	76.16	87.01	78.34
LATex [45]	85.26	77.67	89.40	81.15
SD-ReID [37]	85.16	75.40	85.97	77.02
GSAlign [41]	83.75	75.01	84.10	77.73
GeoReID	**87.02**	**79.46**	**90.64**	**83.55**

TABLE II
PERFORMANCE COMPARISON ON THE AG-REIDV2 DATASET ACROSS FOUR PROTOCOLS. BEST RESULTS ARE HIGHLIGHTED AS **BOLD**, AND SECOND BEST RESULTS ARE HIGHLIGHTED AS UNDERLINED.

Model	A→G		G→A		A→W		W→A	
	Rank1	mAP	Rank1	mAP	Rank1	mAP	Rank1	mAP
Swin [32]	68.76	57.66	68.80	57.70	68.49	56.15	64.40	53.90
HRNet-18 [52]	75.21	65.07	76.25	66.16	76.26	66.17	76.25	66.17
SwinV2 [53]	76.44	66.09	77.11	62.14	80.08	69.09	74.53	65.61
MGN (R50) [54]	82.09	70.17	84.21	72.41	88.14	78.66	84.06	73.73
BoT (R50) [48]	80.73	71.49	79.46	69.67	86.06	75.98	82.69	72.41
SBS (R50) [47]	81.43	72.19	80.15	70.37	86.66	76.68	83.29	73.11
ViT [51]	85.40	77.03	84.65	75.90	89.77	80.48	84.27	76.59
TransReID [15]	88.00	81.40	87.60	80.10	90.40	84.50	87.70	81.10
FusionReID [56]	86.70	80.70	87.90	80.00	89.70	84.20	86.50	80.90
CLIP-ReID [19]	85.36	79.79	85.64	79.08	89.14	84.23	86.50	79.55
PCL-CLIP [57]	79.80	72.20	81.12	72.40	87.14	77.70	84.19	73.89
V2E AGREIDv2 [26]	88.77	80.72	87.86	78.51	93.62	84.85	88.61	80.11
Explain [55]	87.70	79.00	87.35	78.24	**93.67**	83.14	87.73	79.08
VDT [27]	86.46	79.13	86.14	78.12	90.00	82.21	85.26	78.52
SeCap [31]	88.12	80.84	88.24	79.99	91.44	84.01	87.56	80.15
GSAlign [41]	87.86	81.38	88.02	81.05	90.63	83.98	87.31	80.90
SD-ReID [37]	87.04	80.61	86.74	79.24	90.86	84.06	87.86	81.01
LATex [45]	89.13	83.50	89.01	82.85	91.35	86.35	89.32	83.30
GeoReID	**91.26**	**85.52**	**90.34**	**83.88**	93.21	**88.03**	**90.95**	**84.79**

Authorized licensed use limited to: b-on: UNIVERSIDADE DA BEIRA INTERIOR. Downloaded on October 21,2026 at 06:32:31 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Information Forensics and Security. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2026.3744610

The exception is A→W, where Explain [55] and V2E [26] reach marginally higher Rank-1 (93.67% and 93.62% versus 93.21%) but clearly lower mAP (83.14% and 84.85% versus 88.03%), indicating that GeoReID orders the gallery more reliably even when the single top match is occasionally missed. Under the most challenging A→G setting, GeoReID reaches 91.26% Rank-1 accuracy and 85.52% mAP, outperforming all competing methods. Consistent improvements are also observed under G→A, where GeoReID achieves 90.34% Rank-1 accuracy and 83.88% mAP. Notably, GeoReID yields strong gains under the A→W and W→A protocols, which involve low elevation wearable viewpoints and severe viewpoint asymmetry, achieving the highest mAP in both cases. These results demonstrate that rectifying geometry induced distortion in the similarity space leads to more reliable cross view ranking under heterogeneous camera configurations.

TABLE III
PERFORMANCE COMPARISON ON THE CARGO DATASET UNDER METADATA FREE SETTINGS. THE CARGO DATASET DOES NOT PROVIDE CAMERA GEOMETRY ANNOTATIONS. WE GENERATE ALTITUDE, HORIZONTAL DISTANCE, AND VIEWING ANGLE USING OUR VISION ONLY META PREDICTOR. THE SAME PREDICTED GEOMETRY IS PROVIDED TO ALL GEOMETRY AWARE BASELINES, WHILE METHODS WITHOUT GEOMETRY MODELING ARE EVALUATED WITHOUT GEOMETRY INFORMATION. BEST RESULTS ARE HIGHLIGHTED AS **BOLD**, AND SECOND BEST RESULTS ARE HIGHLIGHTED AS UNDERLINED.

Model	All		G→G		A→A		A→G	
	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
SBS [47]	50.32	43.09	72.31	62.99	67.50	49.73	31.25	29.00
PCB [63]	51.00	44.50	74.10	67.60	55.00	44.60	34.40	30.40
BoT [48]	54.81	46.49	77.68	66.47	65.00	49.79	36.25	32.56
MGN [54]	54.81	49.08	83.93	71.05	65.00	52.96	31.87	33.47
VV [50]	45.83	38.84	72.31	62.99	67.50	49.73	31.25	29.00
AGW [60]	60.26	53.44	81.25	71.66	67.50	56.48	43.57	40.90
BAU [61]	45.20	38.40	61.60	51.20	50.00	42.60	40.40	36.70
PAT [62]	37.90	15.30	52.70	24.20	50.00	23.10	35.10	15.50
DTST [43]	64.42	55.73	78.57	72.40	80.00	63.31	50.53	43.49
ViT [51]	61.54	53.54	82.14	71.34	80.00	64.47	43.13	40.11
VDT [27]	64.10	55.20	82.14	71.59	82.50	66.83	48.12	42.76
TransReID [15]	**73.70**	**64.70**	85.70	77.90	**85.00**	**71.80**	64.40	55.90
FusionReID [56]	67.90	61.50	85.70	79.40	80.00	69.30	48.30	53.10
CLIP-ReID [19]	68.27	64.25	84.82	**80.80**	75.00	65.42	55.62	53.83
PCL-CLIP [57]	67.31	60.93	84.82	76.00	70.00	60.75	54.43	51.43
SD-ReID [37]	70.48	61.42	81.25	74.08	82.50	67.70	48.75	46.37
GSAlign [41]	65.06	57.95	83.04	73.86	80.00	65.55	64.89	61.55
SeCap [31]	68.59	60.19	86.61	75.42	80.00	68.08	69.43	58.94
GeoReID	71.79	63.69	**88.04**	79.14	82.51	68.20	**72.02**	**64.61**

TABLE IV
PERFORMANCE COMPARISON ON THE DETREIDX DATASET ACROSS THREE EVALUATION PROTOCOLS. RESULTS ARE REPORTED USING THE BEST VALIDATION CHECKPOINT. BEST RESULTS ARE HIGHLIGHTED AS **BOLD**, AND SECOND BEST RESULTS AS UNDERLINED.

Method	G→A		A→G		A→A	
	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
SBS [47]	46.74	13.53	15.59	15.51	13.74	9.09
MGN [54]	46.48	14.62	15.10	16.33	13.06	8.97
AGW [60]	30.59	8.56	12.30	13.13	10.85	6.84
VDT [27]	50.36	16.37	18.14	18.55	15.27	9.35
SeCap [31]	**51.63**	14.86	17.05	17.77	14.03	9.12
GeoReID	48.87	**21.28**	**20.84**	**23.13**	**16.28**	**12.97**

Quantitative results on the CARGO dataset are reported in Table III. CARGO represents a realistic metadata free aerial-ground ReID scenario, where camera geometry annotations are unavailable and must be estimated. Under the most challenging A→G protocol, GeoReID attains 72.02% Rank-1 accuracy and 64.61% mAP, outperforming all competing methods by a clear margin. This setting involves the largest viewpoint, scale, and perspective discrepancies, highlighting the effectiveness of geometry conditioned similarity rectification for difficult cross view matching. On the aggregate All protocol GeoReID is second to TransReID (71.79% versus 73.70% Rank-1), and performance on the same view protocols (G→G and A→A) remains competitive. This pattern is expected: the All and same view protocols are dominated by ground-to-ground pairs, where geometry conditioning has little to correct, whereas the cross view A→G protocol is where geometry induced distortion is strongest and where GeoReID gains most. These results show that GeoReID remains effective even when camera geometry is predicted rather than provided. Results on the DetReIDX dataset are summarized in Table IV. DetReIDX is particularly challenging due to noisy detections, low image resolution, and extreme aerial-ground viewpoint differences, leading to low absolute performance across all methods. GeoReID achieves the best performance under the A→G protocol, obtaining 20.84% Rank-1 accuracy and 23.13% mAP. Across all three evaluation protocols, GeoReID consistently achieves the highest mAP, indicating more reliable ranking quality under severe geometric distortion. Under the G→A setting, GeoReID attains the highest mAP despite slightly lower Rank-1 accuracy, while performance on the A→A protocol also improves consistently. These results confirm that explicitly rectifying geometry induced similarity distortion is particularly beneficial in extremely challenging cross view scenarios.

TABLE V
ABLATION AND GEOMETRY EMBEDDING ANALYSIS ON DETREIDX

Method	G→A		A→G		A→A	
	R1	mAP	R1	mAP	R1	mAP
Baseline (SeCap)	51.63	14.86	17.05	17.77	14.03	9.12
+ GCPG	46.33	19.47	19.13	21.30	13.73	11.74
+ GIQT	46.59	19.82	19.63	22.04	15.30	12.14
+ GCPG + GIQT	48.87	21.28	20.84	23.13	16.28	12.97
Geometry Embedding Analysis						
w/o Cam	46.57	19.38	18.46	21.04	14.19	11.93
w/o Alt	47.51	20.36	20.11	22.56	14.81	11.77
w/o Ang	47.31	20.15	20.22	22.54	14.94	12.06

TABLE VI
ABLATION STUDY ON AG-REID

Method	A↔G		G↔A	
	R1	mAP	R1	mAP
Baseline (SeCap)	84.03	76.16	87.01	78.34
+ GCPG	86.46	78.68	89.50	82.24
+ GIQT	85.47	77.36	87.97	79.73
+ GCPG + GIQT	87.02	79.46	90.64	83.55
Geometry Embedding Analysis				
w/o Cam	86.37	79.39	89.50	82.61
w/o Alt	84.50	75.99	86.38	78.77
w/o Ang	86.37	78.83	88.88	81.52

D. Ablation Studies and Analysis

Component Ablation Analysis

This section analyzes the contribution of each proposed component and geometry cue across four benchmarks with

Authorized licensed use limited to: b-on: UNIVERSIDADE DA BEIRA INTERIOR. Downloaded on October 21,2026 at 06:32:31 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Information Forensics and Security. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2026.3744610

TABLE VII
ABLATION STUDY ON AG-REIDV2

Method	A↔G		G↔A		A↔W		W↔A	
	R1	mAP	R1	mAP	R1	mAP	R1	mAP
Baseline (SeCap)	88.12	80.84	88.24	79.99	91.44	84.01	87.56	80.15
+ GCPG	90.24	84.41	88.90	83.45	93.30	87.65	90.51	84.21
+ GIQT	88.76	82.59	88.13	81.70	91.13	85.63	88.76	82.59
+ GCPG + GIQT	91.26	85.52	90.34	83.88	93.21	88.03	90.95	84.79
Geometry Embedding Analysis								
w/o Cam	90.32	84.71	89.84	83.54	93.30	87.80	90.94	84.55
w/o Alt	88.67	82.61	88.57	81.76	91.58	85.71	88.85	82.58
w/o Ang	90.15	84.38	89.84	83.65	92.76	87.83	90.43	84.86

TABLE VIII
ABLATION STUDY ON CARGO

Method	All		G→G		A→A		A→G	
	R1	mAP	R1	mAP	R1	mAP	R1	mAP
Baseline (SeCap)	68.59	60.19	86.61	75.42	80.00	68.08	69.43	58.94
+ GCPG	70.83	63.94	86.61	80.79	85.00	69.08	71.28	68.41
+ GIQT	72.76	64.62	88.39	79.40	82.50	68.90	72.34	68.52
+ GCPG + GIQT	71.79	63.69	88.04	79.14	82.51	68.20	72.02	64.61
Geometry Embedding Analysis								
w/o Cam	69.87	63.29	85.71	80.31	82.50	69.55	67.02	65.23
w/o Alt	69.23	63.60	84.82	80.36	85.00	68.50	69.15	68.95
w/o Ang	69.87	63.97	86.61	81.35	80.00	66.74	69.15	67.73

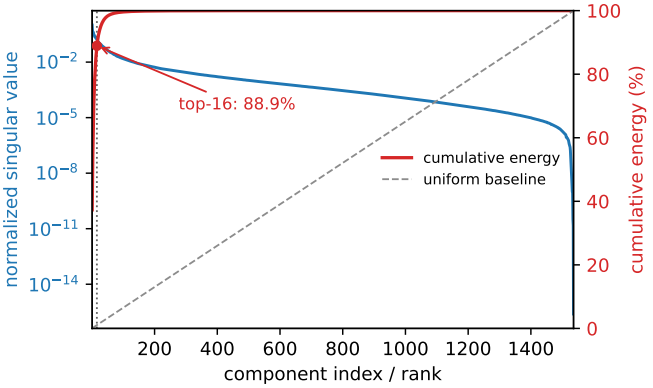


Fig. 4. Structure of the cross view feature covariance difference $\Delta\Sigma$ (d=1536). Singular value spectrum (blue, left axis) and cumulative energy (red, right axis) against a uniform energy baseline (dashed). The top-16 directions capture 88.9% of the distortion energy (marked; top-8: 76.3%), versus only 1.0% under uniformity, demonstrating that geometry induced distortion is highly anisotropic and concentrated in a few dominant directions.

progressively increasing geometric difficulty: AG-ReID, AG-ReIDv2, CARGO, and DetReIDX.

1) Component Level Analysis: We first evaluate the individual and combined effects of geometry conditioned prompts and the GIQT. As shown in Table VI, adding GCPG improves Rank-1 accuracy from 84.03% to 86.46%, confirming that global representation priors benefit from explicit geometry conditioning. GIQT alone provides smaller but consistent gains, indicating the presence of local similarity distortion even when global features are unchanged. Combining GCPG and GIQT yields the best results across both A↔G and G↔A protocols, demonstrating complementary effects between global adaptation and local similarity rectification. Table VII shows that the same trend holds under more diverse camera configurations, including wearable views. GCPG provides strong improvements across all protocols, while GIQT yields consistent gains, particularly in cross view settings. The combined GCPG+GIQT model achieves the best overall performance, with notable improvements under the challenging A↔G and W↔A protocols, where viewpoint asymmetry is most severe. CARGO represents a metadata free scenario where geometry must be predicted. As shown in Table VIII, GIQT yields larger gains than GCPG under the most challenging A→G protocol, raising Rank-1 accuracy from the 69.43% baseline to 72.34%, compared with 71.28% for GCPG. This suggests that under extreme cross view mismatch and imperfect geometry estimation, correcting similarity space distortion becomes more critical than global prompt adaptation alone. DetReIDX is characterized by noisy detections, low resolution, and extreme aerial-ground viewpoint differences. As shown in Table V, GCPG and GIQT individually provide modest gains, while their combination achieves the highest mAP across all protocols. Although absolute performance remains low due to dataset difficulty, the consistent improvement in mAP indicates more reliable ranking under severe geometric distortion.

Geometry Embedding Analysis

We further analyze (Table V, VI, VII, VIII) the contribution of individual geometry cues by removing camera identity, altitude, and viewing angle encoding. Across all datasets, removing altitude or viewing angle information leads to larger performance degradation than removing camera identity. For example, on AG-ReID (Table VI), removing altitude reduces A↔G Rank-1 accuracy from 87.02% to 84.50%, while removing camera identity results in a smaller drop. Similar trends are observed on AG-ReIDv2, CARGO, and DetReIDX. These results indicate that geometry cues altitude and viewing angle are the dominant factors governing geometry induced similarity distortion, while camera identity primarily provides coarse contextual bias.

E. Analysis of Design Choices

This subsection provides empirical justification for the architectural and modeling choices of the proposed framework. All analyzes are conducted consistently across AG-ReID, AG-ReIDv2, CARGO, and DetReIDX, covering both metadata rich and metadata free settings, as well as moderate to extreme viewpoint disparity.

Structure of Geometry Induced Distortion

We analyze the structure of cross view geometric distortion through the covariance difference $\Delta\Sigma = \Sigma_{\text{aerial}} - \Sigma_{\text{ground}}$ of the final ReID descriptors (d=1536), with eigenvalues $\{\lambda_i\}_{i=1}^{d}$. To quantify anisotropy rather than argue it qualitatively, we report three scale invariant concentration measures: the normalized *participation ratio* $\text{PR} = \big(\sum_i \lambda_i\big)^2 / \big(d \sum_i \lambda_i^2\big) \in (0, 1]$, the *effective rank* $r_{\text{eff}} = \exp\big(-\sum_i p_i \log p_i\big)$ with $p_i = \lambda_i / \sum_j \lambda_j$, and the *stable rank* $\|\Delta\Sigma\|_F^2 / \|\Delta\Sigma\|_2^2$. Under perfect isotropy these attain their maxima (PR=1, r_{eff}=d); values far below the maximum indicate concentration in a few directions. As shown in Fig. 4, the spectrum exhibits pronounced decay: the top-8 components capture 76.3% of the total energy and the top-16 capture 88.9%, whereas an isotropic spectrum would capture only $8/1536 \approx 0.5\%$ and 1.0% respectively. Correspondingly, the

Authorized licensed use limited to: b-on: UNIVERSIDADE DA BEIRA INTERIOR. Downloaded on October 21,2026 at 06:32:31 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Information Forensics and Security. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2026.3744610

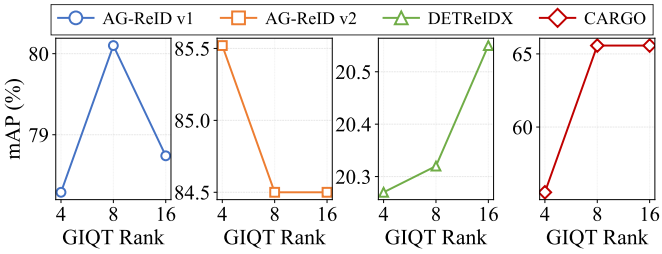


Fig. 5. Performance sensitivity with respect to GIQT rank across datasets. Low rank correction (8-16) yields optimal performance.

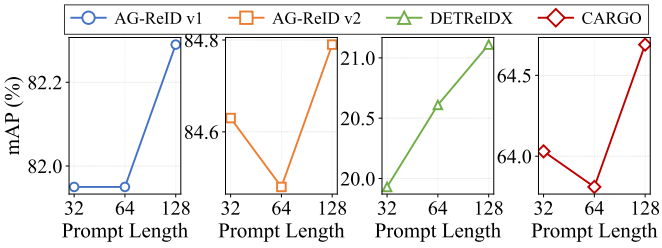


Fig. 6. Effect of geometry conditioned prompt length across datasets. Performance remains stable over a wide range of prompt sizes.

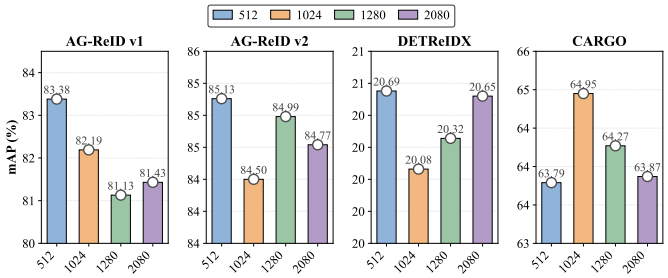


Fig. 7. Effect of hidden dimension size across datasets. Increasing feature dimensionality beyond 512 does not yield consistent gains, indicating that geometry induced similarity distortion not model capacity is the dominant bottleneck.

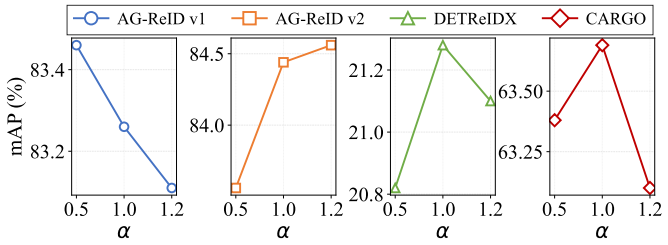


Fig. 8. Sensitivity to geometry conditioned prompt scaling factor α across datasets. Performance peaks near $\alpha = 1.0$, indicating that geometry prompts should bias representations rather than dominate identity features.

measured participation ratio (PR $= 0.030$), effective rank ($r_{\text{eff}} = 121.7$), and especially the stable rank (2.68) are far below the isotropic maxima, indicating that the distortion energy is dominated by fewer than three directions. To confirm this is not a finite sample artefact, we compare against a label permutation null in which the aerial-ground assignments are randomly shuffled, so that under the null no geometry induced structure exists: over 1000 permutations the observed top-16 energy exceeds every null sample (permutation $p < 0.001$). This statistically establishes that geometry induced distortion is anisotropic and concentrated in a small number of dominant directions, directly motivating the low rank residual formulation of GIQT, which targets these directions without overfitting higher-order variation.

Sensitivity to GIQT Rank

Figure 5 analyzes the effect of the GIQT rank on performance across AG-ReIDv1, AG-ReIDv2, DetReIDX, and CARGO. Across all datasets, performance improves when increasing the rank from 4 to 8 and either saturates or slightly degrades beyond rank 16. In particular, AG-ReIDv1 achieves its peak mAP at rank 8, while AG-ReIDv2 and CARGO exhibit marginal gains up to rank 16. DetReIDX, which features more severe geometric distortion, also benefits most from low to moderate ranks. These results indicate that geometry induced distortion is dominated by a small number of directions and that low rank correction (8-16) is sufficient, while higher ranks provide diminishing returns and risk of overfitting.

Sensitivity to geometry conditioned prompt length

Figure 6 evaluates the impact of geometry conditioned prompt length. Performance remains stable across a wide range of prompt sizes (32-128 tokens) on all datasets. AG-ReIDv1 and AG-ReIDv2 show minor fluctuations, while DetReIDX and CARGO exhibit slightly better performance with longer prompts. These results demonstrate that the proposed GCPG mechanism is robust to prompt length and does not require careful tuning, indicating that geometry information is encoded compactly rather than relying on excessive prompt capacity.

Hidden Dimension Sensitivity Figure 7 analyzes the effect of increasing the hidden dimension across all four benchmarks. Across AG-ReID, AG-ReIDv2, CARGO, and DetReIDX, increasing the hidden dimension beyond 512 does not yield consistent improvements and in several cases slightly degrades performance. This behavior indicates that cross view performance degradation under extreme aerial-ground geometry is not caused by insufficient representational capacity. Simply increasing feature dimensionality does not resolve the dominant failure mode. Instead, the limiting factor lies in the mismatch of the similarity space induced by geometric distortion. This observation provides further justification for explicitly rectifying similarity computation via geometry conditioned query-key transformation, rather than increasing model capacity or relying on wider embeddings.

Sensitivity to Prompt Scaling Figure 8 analyzes the effect of the geometry conditioned prompt scaling factor α across all four benchmarks. Across datasets, performance consistently peaks around $\alpha = 1.0$ and degrades when the prompt contribution is either underscaled or over emphasized. When α is too small, geometry conditioned prompts fail to provide sufficient bias to compensate for viewpoint induced distortion. Conversely, when α is too large, geometry cues begin to

Authorized licensed use limited to: b-on: UNIVERSIDADE DA BEIRA INTERIOR. Downloaded on October 21,2026 at 06:32:31 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Information Forensics and Security. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2026.3744610

TABLE IX

ROBUSTNESS TO DISCRETIZED GEOMETRY METADATA CORRUPTION ACROSS DATASETS. RESULTS ARE REPORTED UNDER PROTOCOL 1 (A↔G) FOR AG-REID AND A→G FOR DETREIDX. AG-REID USES 3 ALTITUDE BINS AND 3 VIEWING ANGLE BINS, WHILE DETREIDX USES 6 ALTITUDE BINS AND 3 VIEWING ANGLE BINS. ALL CORRUPTIONS ARE APPLIED *only at inference time* TO GEOMETRY INPUTS, USING MODELS TRAINED WITH CLEAN GEOMETRY. ΔR1 (BASELINE) IS COMPUTED RELATIVE TO THE GEOMETRY AGNOSTIC SECAP BASELINE (84.03% RANK-1 FOR AG-REID AND 17.05% RANK-1 FOR DETREIDX).

Corruption Type	AG-ReID		DetReIDX	
	Rank-1	ΔR1 (Base)	Rank-1	ΔR1 (Base)
Clean geometry	87.02	+2.99	20.84	+3.79
Random bin flip (Altitude)	86.12	+2.09	20.12	+3.07
Random bin flip (Angle)	85.71	+1.68	19.63	+2.58
Joint bin perturbation	84.98	+0.95	19.05	+2.00
Biased altitude shift	84.61	+0.58	18.67	+1.62
Stale geometry	85.23	+1.20	19.34	+2.29
Wrong geometry	84.17	+0.14	18.21	+1.16
Coarse geometry bins	86.45	+2.42	20.38	+3.33
Missing/default geometry	85.06	+1.03	19.02	+1.97

TABLE X

ZERO SHOT UNSEEN GEOMETRY GENERALIZATION UNDER LEAVE ONE BIN OUT EVALUATION ON AG-REID. TO EVALUATE WHETHER THE MODEL GENERALIZES TO GEOMETRY CONFIGURATIONS NOT OBSERVED DURING TRAINING, ONE ALTITUDE BIN IS ENTIRELY REMOVED FROM THE TRAINING SET AND USED ONLY AT TEST TIME. *Seen* REPORTS PERFORMANCE ON QUERY IMAGES FROM THE RETAINED ALTITUDE BINS, WHILE *Unseen* REPORTS PERFORMANCE ON QUERY IMAGES FROM THE HELD OUT ALTITUDE BIN. FOR GEOREID, THE UNSEEN GEOMETRY BIN IS MAPPED EITHER TO THE NEAREST RETAINED BIN (*nearest*) OR TO THE DEFAULT GEOMETRY EMBEDDING (*default*). ΔUNSEEN DENOTES THE RANK-1 GAIN OVER THE GEOMETRY AGNOSTIC SECAP BASELINE ON THE SAME HELD OUT BIN.

Method	A→G			G→A		
	Seen R1	Unseen R1	ΔUnseen	Seen R1	Unseen R1	ΔUnseen
SeCap	83.27	80.56	0.00	86.09	82.19	0.00
GeoReID (nearest)	**86.12**	**83.62**	**+3.06**	**88.29**	**85.47**	**+3.28**
GeoReID (default)	85.34	82.22	+1.66	87.69	84.39	+2.20

dominate identity discriminative features, leading to reduced retrieval performance. This behavior is particularly evident on CARGO and DetReIDX, where extreme viewpoint disparity and noisy geometry estimation amplify the negative effects of over scaling. These results confirm that geometry conditioned prompts should act as a structured bias rather than a dominant representation component. The residual prompt formulation enables controlled adaptation, allowing geometry information to guide feature extraction without overwhelming identity semantics.

Corruption definitions.

Random bin flip (Altitude / Angle) randomly shifts the corresponding geometry bin by ± 1, clipped to valid bin boundaries. *Joint bin perturbation* applies independent ± 1 shifts to both altitude and angle bins. *Biased altitude shift* applies a fixed $+1$ altitude bin offset to all samples. *Stale geometry* assigns geometry bins from a temporally adjacent or nearby sample, simulating delayed or asynchronous metadata. *Wrong geometry* randomly shuffles geometry bins across identities, breaking geometry image correspondence while preserving the marginal bin distribution. All perturbations are applied independently per image and consistently to both query and gallery samples.



Fig. 9. Robustness to discretized geometry metadata corruption. The y-axis shows Rank-1 degradation relative to clean geometry. All corruptions are applied only at inference time using models trained with clean geometry. Results are reported for AG-ReID (A↔G) and DetReIDX (A→G).

Table IX and Fig. 9 summarize robustness to inference time geometry corruption. Across all perturbations GeoReID stays above the geometry agnostic SeCap baseline: correct geometry perturbations (bin flips, joint and biased shifts) cause only graceful degradation, degraded or coarsened metadata (stale, coarse bins) is tolerated with a small margin, and even under *Wrong geometry* performance falls back to approximately the baseline ($+0.14$ Rank-1) rather than below it. This shows geometry is exploited as a helpful bias without a catastrophic dependency.

Zero shot unseen geometry generalization. The corruption study operates at inference time on models trained with clean geometry. To test generalization to geometry *never seen during training*, we adopt a leave one bin out protocol: one altitude bin is entirely excluded from training and used only at test time. Table X reports performance on the held out (unseen) bin and the retained (seen) bins. GeoReID retains a clear advantage over the geometry agnostic baseline on the unseen bin ($+3.06$ Rank-1 under the nearest bin policy), and its seen→unseen drop is comparable to the baseline's own drop, indicating that geometry conditioning generalizes rather than memorizing the training geometry distribution.

F. GIQT Placement and Distinction from Conditional Attention

To further clarify the architectural role of GIQT, we ablate its placement inside the CVFT decoder in Table XI. Specifically, we compare applying the geometry conditioned query-key transformation to self attention versus cross attention, while keeping all other components unchanged. The results show that cross attention placement is more effective, supporting our design choice that geometry aware rectification should be applied during promptto image cross view matching rather than intra token self refinement.

Beyond the empirical placement result, Table XII contrasts GIQT with related conditioning mechanisms. Feature modulation and FiLM style methods scale or shift feature *channels* and thus alter token content; conditional attention variants reweight tokens or reparametrize projections; geospatial attention [68] conditions *where* to attend using calibrated geo cues. GIQT is distinct in that it conditions the query-key comparison *metric* on camera geometry while leaving the value stream untouched: it changes *how* tokens are compared rather than what they contain or which are emphasized.

Authorized licensed use limited to: b-on: UNIVERSIDADE DA BEIRA INTERIOR. Downloaded on October 21,2026 at 06:32:31 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Information Forensics and Security. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2026.3744610

TABLE XI

ABLATION OF GIQT PLACEMENT WITHIN THE CVFT DECODER ON AG-REIDV1. BOTH VARIANTS KEEP THE SAME GCPG PROMPTS, LOSSES, GEOMETRY ENCODING, AND TRAINING PROTOCOL; ONLY THE ATTENTION LAYER RECEIVING THE GEOMETRY CONDITIONED QUERY-KEY TRANSFORMATION IS CHANGED. THE COMPARISON ISOLATES WHETHER GEOMETRY AWARE RECTIFICATION IS MORE EFFECTIVE IN INTRA TOKEN SELF ATTENTION OR IN CROSS ATTENTION BETWEEN PROMPT TOKENS AND IMAGE TOKENS.

GIQT Placement	A→G		G→A	
	Rank-1	mAP	Rank-1	mAP
Self-attention only	84.71	76.94	87.85	80.57
Cross-attention only	**87.02**	**79.46**	**90.64**	**83.55**

TABLE XII

DISTINCTION FROM CONDITIONAL AND GEOMETRY AWARE ATTENTION. GIQT CONDITIONS THE QUERY-KEY SIMILARITY METRIC ON CAMERA GEOMETRY WHILE KEEPING THE VALUE STREAM UNCHANGED.

Mechanism	Signal	Acts on	Effect
Feature modulation / FiLM	attributes/context	feature channels	changes content
Conditional attention	image/context	weights/projections	reweights tokens
Geospatial attention [68]	calibrated geo cues	spatial attention	changes where to attend
GIQT	camera geometry	Q-K metric	changes how tokens compare

TABLE XIII

COMPUTATIONAL COST ANALYSIS AT 256×128 INPUT. LATENCY MEASURED ON ONE NVIDIA A40. EACH METHOD USES ITS OFFICIAL PATCH STRIDE. THE PROPOSED MODULES ADD +0.27% FLOPS AND ∼0.01 MS PER IMAGE OVER SECAP WITHIN MEASUREMENT NOISE AS GIQT IS LOW RANK (r=4) AND GCPG EXECUTES ONCE PER IMAGE ON A SINGLE DESCRIPTOR RATHER THAN OVER THE TOKEN SEQUENCE.

Method	Params (M)	FLOPs (G)	ms/img	img/s
VDT [27]	85.8	22.9	1.05	955
SeCap [31]	130.9	46.7	2.13	469
Ours (GIQT only)	135.8	46.7	2.13	469
Ours (full)	188.1	46.8	2.14	467

TABLE XIV

GEOMETRY PREDICTOR ACCURACY UNDER DATASET SPECIFIC GEOMETRY LABEL AVAILABILITY. DETREIDX IS EVALUATED ON IDENTITY DISJOINT HELD OUT TEST PIDS. AG-REID PROVIDES ONLY ALTITUDE/HEIGHT METADATA; ORIGINAL DISTANCE AND VIEWING ANGLE LABELS ARE UNAVAILABLE.

Dataset	Altitude MAE	Distance MAE	Angle Acc.
DetReIDX	2.726 m	3.148 m	94.31%
AG-ReID	1.87 m	N/A	N/A

G. Computational Overhead

We quantify the cost of the proposed modules in Table XIII, reporting parameters, FLOPs, per image latency and throughput at 256×128 input on a single NVIDIA A40. Relative to the SeCap baseline, the full GeoReID model adds only $+0.27\%$ FLOPs ($46.7\rightarrow46.8$ G) and ~0.01 ms per image ($2.13\rightarrow2.14$ ms), leaving throughput essentially unchanged ($469\rightarrow467$ img/s). This is because GIQT is a low rank (r=4) per head transform applied only to queries and keys, and GCPG is evaluated once per image on a single global descriptor rather than over the full token sequence. The parameter increase does not translate into a meaningful compute or latency cost, supporting the "minimal overhead" claim relevant to UAV deployment. The FLOPs gap between VDT and SeCap stems from the finer patch stride (12 vs 16), which our method inherits unchanged.

Table XIV additionally reports the accuracy of the vision only geometry predictor used when metadata is unavailable. On DetReIDX (evaluated on identity disjoint held out PIDs) it attains 2.73 m altitude MAE, 3.15 m distance MAE and 94.31% viewing angle accuracy; on AG-ReID, where only altitude/height is annotated, it reaches 1.87 m altitude MAE. These errors are small relative to the operating altitude range (15–45 m) and, as shown by the corruption analysis (Table IX), well within the range GeoReID tolerates, explaining why predicted geometry remains effective on the metadata free CARGO benchmark.

H. Visualization Analysis

Qualitative Retrieval Analysis

Figure 10 presents representative Rank-1 to Rank-5 retrieval results comparing the baseline (SeCap) and the proposed method under challenging aerial–ground matching scenarios. Correct matches are highlighted in green, while incorrect matches are shown in red. The baseline frequently retrieves visually similar but incorrect identities under extreme viewpoint changes, particularly when aerial queries induce strong foreshortening, scale compression, or body part displacement. In these cases, similarity scores are dominated by spurious local alignments, leading to incorrect high rank matches despite overall appearance resemblance. In contrast, the proposed method consistently retrieves correct identities at higher ranks. Notably, improvements are observed not only at Rank-1 but also across the entire top-k list, indicating more reliable similarity ordering rather than isolated corrections. This behavior suggests that geometry induced distortion in the similarity space is explicitly mitigated, enabling more consistent alignment of corresponding body regions across views. These qualitative results complement the quantitative gains, illustrating that geometry conditioned similarity rectification improves ranking stability under extreme cross view viewpoint discrepancies.

I. Feature Space Visualization

Figure 11 provides the learned feature distributions of the baseline model and the proposed method using t-SNE. Each color denotes a identity, while circles and crosses represent aerial and ground views, respectively. In the baseline model, features corresponding to the same identity but different views exhibit noticeable separation and overlap with other identities. This indicates that although identity level clustering exists, cross view alignment remains inconsistent, leading to fragmented clusters and view dependent feature drift. In contrast, the proposed method produces more compact and well aligned clusters across views. Features from aerial and ground cameras corresponding to the same identity are more tightly grouped, while inter identity separation is preserved. Importantly, this

Authorized licensed use limited to: b-on: UNIVERSIDADE DA BEIRA INTERIOR. Downloaded on October 21,2026 at 06:32:31 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Information Forensics and Security. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2026.3744610



Fig. 10. Qualitative comparison of Rank-1 to Rank-5 retrieval results between the baseline (SeCap) and the proposed method. The baseline suffers from incorrect high rank matches due to geometry induced similarity distortion, while the proposed method produces more consistent and correct ranking under extreme aerial-ground viewpoint differences.

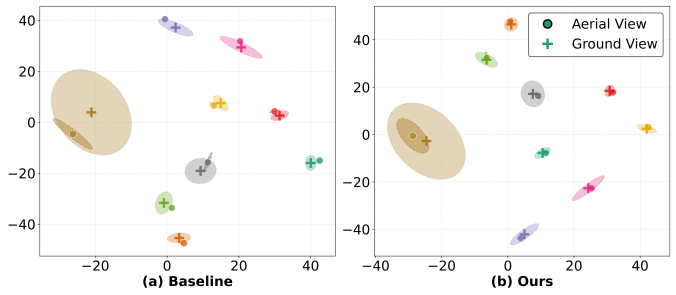


Fig. 11. 2D t-SNE visualization of aerial and ground view feature embeddings for the baseline and our proposed method. Circles (●) represent the geometric mean aerial view samples, and plus symbols (+) represent ground view samples, while the shaded areas provide the ellipses that contain 90% of the corresponding data points. Colors indicate identity labels.

improvement is not achieved by collapsing representations, but by reducing cross view dispersion within each identity cluster. While t-SNE is a qualitative visualization tool, the observed reduction in cross view overlap and improved cluster compactness are consistent with the quantitative gains reported earlier. These results suggest that geometry conditioned prompt adaptation and similarity space rectification jointly encourage more view consistent representations.

V. Conclusion

This work identified the *geometry induced distortion in the similarity space* factor as a key failure mode in aerial-ground person re-identification. As a consequence, we concluded that the geometry invariant similarity assumption underlying attention does not hold, even when feature representations are partially aligned. To address this issue, we introduced a geometry conditioned similarity alignment framework that explicitly incorporates camera geometry into both global representation adaptation and local similarity computation. The proposed Geometry Induced Query-Key Transformation (GIQT) rectifies dominant, anisotropic similarity distortions through a lightweight low rank formulation, while geometry conditioned prompts provide complementary global adaptation. Extensive experiments across four benchmarks suggest that explicitly correcting the similarity space consistently improves robustness under extreme, corrupted, and unseen geometric conditions, with minimal computational overhead. These findings highlight the importance of geometry aware similarity modeling for reliable cross view recognition in *real-world* aerial-ground scenarios.

Acknowledgements

This work is funded by national funds through FCT – Fundação para a Ciência e a Tecnologia, I.P., and, when eligible, co-funded by EU funds under project/support UID/50008/2025 – Instituto de Telecomunicações, with DOI identifier https://doi.org/10.54499/UID/50008/2025.

References

[1] S. Zhang, Q. Yang, D. Cheng, Y. Xing, G. Liang, P. Wang, and Y. Zhang, "Ground-to-Aerial Person Search: Benchmark Dataset and Approach," *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 789–799, 2023.

[2] Y.-C. Chen, X. Zhu, W.-S. Zheng, and J.-H. Lai, "Person re-identification by camera correlation aware feature augmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 2, pp. 392–408, Feb. 2018.

[3] L. Zheng, Y. Yang, and A. G. Hauptmann, "Person re-identification: Past, present and future," arXiv:1610.02984, 2016.

[4] X. Liu, W. Liu, H. Ma, and H. Fu, "Large-scale vehicle re-identification in urban surveillance videos," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2016, pp. 1–6.

[5] X. Liu, W. Liu, T. Mei, and H. Ma, "A deep learning-based approach to progressive vehicle re-identification for urban surveillance," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 869–884.

[6] H. Li, M. Ye, C. Wang, and B. Du, "Pyramidal transformer with conv-patchify for person re-identification," in *Proc. ACM international conference Multimedia (ACM MM)*, 2022, pp. 7317–7326.

[7] S. Yang, B. Kang, and Y. Lee, "Sampling agnostic feature representation for long-term person re-identification," *IEEE Transactions on Image Processing*, vol. 31, pp. 6412–6423, 2022.

[8] G. Chen, T. Gu, J. Lu, J.-A. Bao, and J. Zhou, "Person re-identification via attention pyramid," *IEEE Transactions on Image Processing*, vol. 30, pp. 7663–7676, 2021.

[9] X. Tian, J. Liu, Z. Zhang, C. Wang, Y. Qu, Y. Xie, and L. Ma, "Hierarchical walking transformer for object re-identification," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 4224–4232.

[10] S. Zhang, Q. Zhang, Y. Yang, X. Wei, P. Wang, B. Jiao, and Y. Zhang, "Person re-identification in aerial imagery," *IEEE Transactions on Multimedia*, vol. 23, pp. 281–291, 2020.

[11] T. Li, J. Liu, W. Zhang, Y. Ni, W. Wang, and Z. Li, "UAV-Human: A large benchmark for human behavior understanding with unmanned aerial vehicles," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 16266–16275.

[12] P. Wang, B. Jiao, L. Yang, Y. Yang, S. Zhang, W. Wei, and Y. Zhang, "Vehicle re-identification in aerial imagery: Dataset and approach," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 460–469.

[13] S. V. A. Kumar, E. Yaghoubi, A. Das, B. S. Harish, and H. Proença, "The P-DESTRE: A fully annotated dataset for pedestrian detection, tracking, and short/long-term re-identification from aerial devices," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 1696–1708, 2020.

Authorized licensed use limited to: b-on: UNIVERSIDADE DA BEIRA INTERIOR. Downloaded on October 21,2026 at 06:32:31 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Information Forensics and Security. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2026.3744610

[14] S. Teng, S. Zhang, Q. Huang, and N. Sebe, "Viewpoint and scale consistency reinforcement for UAV vehicle re-identification," *International Journal of Computer Vision*, vol. 129, no. 3, pp. 719–735, 2021.

[15] S. He, H. Luo, P. Wang, F. Wang, H. Li, and W. Jiang, "TransReID: Transformer-based object re-identification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15013–15022.

[16] W. Li, C. Zou, M. Wang, F. Xu, J. Zhao, R. Zheng, Y. Cheng, and W. Chu, "DC-Former: Diverse and compact transformer for person re-identification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 2, 2023, pp. 1415–1423.

[17] L. Tan, P. Dai, J. Chen, L. Cao, Y. Wu, and R. Ji, "PartFormer: Awakening latent diverse representation from vision transformer for object re-identification," *arXiv preprint arXiv:2408.16684*, 2024.

[18] G. Zhang, Y. Zhang, T. Zhang, B. Li, and S. Pu, "PHA: Patch-wise high-frequency augmentation for transformer-based person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 14133–14142.

[19] S. Li, L. Sun, and Q. Li, "CLIP-ReID: Exploiting vision-language model for image re-identification without concrete text labels," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 1, 2023, pp. 1405–1413.

[20] L. Tan, Y. Zhang, K. Han, P. Dai, Y. Zhang, Y. Wu, and R. Ji, "RLE: A unified perspective of data augmentation for cross-spectral re-identification," *Advances in Neural Information Processing Systems*, vol. 37, pp. 126977–126996, 2024.

[21] Y. Li, J. He, T. Zhang, X. Liu, Y. Zhang, and F. Wu, "Diverse part discovery: Occluded person re-identification with part-aware transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2898–2907.

[22] C. Wang, Y. Yang, M. Qi, H. Zhang, and H. Ma, "Towards Efficient Object Re-Identification with a Novel Cloud-Edge Collaborative Framework," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 7, 2025, pp. 7600–7608.

[23] Y. Feng, J. Li, C. Xie, L. Tan, and J. Ji, "Multi-Modal Object Re-identification via Sparse Mixture-of-Experts," in *Proceedings of the Forty-Second International Conference on Machine Learning*.

[24] Y. Jia, R. Quan, Y. Feng, H. Chen, and J. Qin, "Doubly Contrastive Learning for Source-Free Domain Adaptive Person Search," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 4, 2025, pp. 3949–3957.

[25] H. Nguyen, K. Nguyen, S. Sridharan, and C. Fookes, "Aerial-ground person re-ID," in *2023 IEEE International Conference on Multimedia and Expo (ICME)*, 2023, pp. 2585–2590.

[26] H. Nguyen, K. Nguyen, S. Sridharan, and C. Fookes, "AG-ReID.v2: Bridging aerial and ground views for person re-identification," *IEEE Transactions on Information Forensics and Security*, vol. 19, pp. 2896–2908, 2024.

[27] Q. Zhang, L. Wang, V. M. Patel, X. Xie, and J. 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*, 2024, pp. 22000–22009.

[28] K. A. Hambarde, N. Mbongo, P. K. MP, S. Mekewad, C. Fernandes, G. Silahtaroğlu, A. Nithya, P. Wasnik, M. D. Rashidunnabi, P. Samale, *et al.*, "DetReIDX: A stress-test dataset for real-world UAV-based person recognition," *arXiv preprint arXiv:2505.04793*, 2025.

[29] F. Liu, L. Yao, C. Zhang, T. Wu, X. Zhang, X. Jiang, and J. Zhou, "Boost UAV-based object detection via scale-invariant feature disentanglement and adversarial learning," *IEEE Transactions on Geoscience and Remote Sensing*, 2025.

[30] C. Yan, G. Pang, L. Wang, J. Jiao, X. Feng, C. Shen, and J. Li, "BV-Person: A large-scale dataset for bird-view person re-identification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10943–10952.

[31] S. Wang, Y. Wang, R. Wu, B. Jiao, W. Wang, and P. Wang, "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*, 2025, pp. 22119–22128.

[32] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin Transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10012–10022.

[33] K. Hambarde and H. Proenca, "Image-based human re-identification: Which covariates are actually (the most) important?" *Image and Vision Computing*, vol. 143, p. 104917, 2024.

[34] S. Zhang, W. Luo, D. Cheng, Q. Yang, L. Ran, Y. Xing, and Y. Zhang, "Cross-platform video person re-identification: A new benchmark dataset and adaptation approach," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024, pp. 270–287.

[35] K. Nguyen, F. Liu, C. Fookes, S. Sridharan, X. Liu, and A. Ross, "Person Recognition in Aerial Surveillance: A Decade Survey," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2025.

[36] W. Khalid, B. Liu, X. Li, M. Waqas, and M. S. 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 (ICCV)*, 2025, pp. 9749–9758.

[37] Y. Wang, X. Hu, L. Wang, P. Zhang, and H. Lu, "SD-ReID: View-aware Stable Diffusion for Aerial-Ground Person Re-Identification," *arXiv preprint arXiv:2504.09549*, 2025.

[38] L. Wang, Q. Zhang, J. Qiu, and J. Lai, "Rotation Exploration Transformer for Aerial Person Re-Identification," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, 2024, pp. 1–6.

[39] F. Zhang, E. Firkat, H. Ma, J. Zhu, B. Zhu, and A. Hamdulla, "DARI: Transformer-Based Data Augmentation and Rotation Invariance for UAV Person Re-Identification," *IEEE Transactions on Multimedia*, 2025.

[40] S. Chen, M. Ye, Y. Huang, and B. Du, "Towards Effective Rotation Generalization in UAV Object Re-Identification," *IEEE Transactions on Information Forensics and Security*, 2025.

[41] Q. Li, J. Li, Y. Zhang, L. Tan, J. Chen, and J. Ji, "GSAlign: Geometric and Semantic Alignment Network for Aerial-Ground Person Re-Identification," *arXiv preprint arXiv:2510.22268*, 2025.

[42] J. Qiu, Z. Feng, L. Wang, and J. Lai, "Salient Part-Aligned and Keypoint Disentangling Transformer for Person Re-Identification in Aerial Imagery," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, 2024, pp. 1–6.

[43] Y. Wang and M. Pishgar, "Dynamic Token Selection for Aerial-Ground Person Re-Identification," *arXiv preprint arXiv:2412.00433*, 2024.

[44] Y. Wang, M. Yang, R. Cao, and G. Gao, "AEA-FIRM: Adaptive Elastic Alignment with Fine-Grained Representation Mining for Text-based Aerial Pedestrian Retrieval," *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

[45] P. Zhang, X. Hu, Y. Wang, and H. Lu, "Latex: Leveraging Attribute-Based Text Knowledge for Aerial-Ground Person Re-Identification," *arXiv preprint arXiv:2503.23722*, 2025.

[46] J. Li, Z. Chen, L. Deng, C. Wang, and G. Zhai, "MMReID-Bench: Unleashing the Power of MLLMs for Effective and Versatile Person Re-Identification," *arXiv e-prints*, arXiv:2508, 2025.

[47] L. He, X. Liao, W. Liu, X. Liu, P. Cheng, and T. Mei, "FastReID: A PyTorch Toolbox for General Instance Re-Identification," in *Proceedings of the 31st ACM International Conference on Multimedia (ACM MM)*, 2023, pp. 9664–9667.

[48] H. Luo, Y. Gu, X. Liao, S. Lai, and W. 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 (CVPRW)*, Jun. 2019, pp. 1487–1495.

[49] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang, "Learning Generalisable Omni-Scale Representations for Person Re-Identification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 9, pp. 5056–5069, 2021.

[50] R. Kuma, E. Weill, F. Aghdasi, and P. Sriram, "Vehicle Re-Identification: An Efficient Baseline Using Triplet Embedding," in *2019 International Joint Conference on Neural Networks (IJCNN)*, 2019, pp. 1–9.

[51] A. Dosovitskiy, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *arXiv preprint arXiv:2010.11929*, 2020.

[52] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 10, pp. 3349–3364, 2020.

[53] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, *et al.*, "Swin Transformer V2: Scaling up capacity and resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12009–12019, 2022.

[54] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou, "Learning discriminative features with multiple granularities for person re-identification," in *Proceedings of the 26th ACM International Conference on Multimedia (ACM MM)*, pp. 274–282, 2018.

[55] K. Nguyen, C. Fookes, S. Sridharan, F. Liu, X. Liu, A. Ross, D. Michalski, H. Nguyen, D. Deb, M. Kothari, *et al.*, "AG-ReID 2023: Aerial-ground person re-identification challenge results," in *2023 IEEE International Joint Conference on Biometrics (IJCB)*, pp. 1–10, 2023.

Authorized licensed use limited to: b-on: UNIVERSIDADE DA BEIRA INTERIOR. Downloaded on October 21,2026 at 06:32:31 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted,
but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Information Forensics and Security. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2026.3744610

[56] Y. Wang, P. Zhang, X. Liu, Z. Tu, and H. Lu, "Unity is strength: Unifying convolutional and transformeral features for better person re-identification," *IEEE Transactions on Intelligent Transportation Systems*, 2025.

[57] J. Li and X. Gong, "Prototypical contrastive learning-based CLIP fine-tuning for object re-identification," *arXiv preprint arXiv:2310.17218*, 2023.

[58] D. Mao, S. Teng, and X. Lv, "CVAF: A CLIP-Based View-Consistent Alignment Framework for Aerial-Ground Person Re-Identification," *ACM Transactions on Multimedia Computing, Communications and Applications*, 2025.

[59] T. Wang, H. Liu, P. Song, T. Guo, and W. Shi, "Pose-guided feature disentangling for occluded person re-identification based on transformer," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 3, pp. 2540–2549, 2022.

[60] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. Hoi, "Deep learning for person re-identification: A survey and outlook," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 6, pp. 2872–2893, 2021.

[61] Y. Cho, J. Kim, W. J. Kim, J. Jung, and S.-E. Yoon, "Generalizable person re-identification via balancing alignment and uniformity," *Advances in Neural Information Processing Systems*, vol. 37, pp. 47069–47093, 2024.

[62] H. Ni, Y. Li, L. Gao, H. T. Shen, and J. Song, "Part-aware transformer for generalizable person re-identification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11280–11289, 2023.

[63] Y. Sun, L. Zheng, Y. Li, Y. Yang, Q. Tian, and S. Wang, "Learning part-based convolutional features for person re-identification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 3, pp. 902–917, 2019.

[64] H. Moon and P. J. Phillips, "Computational and performance aspects of PCA-based face-recognition algorithms," *Perception*, vol. 30, no. 3, pp. 303–321, 2001.

[65] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1116–1124.

[66] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proceedings of the 19th International Conference on Computational Statistics (COMPSTAT)*, 2010, pp. 177–186.

[67] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 7, 2020, pp. 13001–13008.

[68] S. Workman, M. U. Rafique, H. Blanton, and N. Jacobs, "Revisiting near/remote sensing with geospatial attention," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 1778–1787.

Authorized licensed use limited to: b-on: UNIVERSIDADE DA BEIRA INTERIOR. Downloaded on October 21,2026 at 06:32:31 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted,
but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.