OSFAL: ORTHOGONAL CROSS-SPECTRAL FEATURE ALIGNMENT AND LAYER-LOCALIZED CROSS-VIEW LEARNING FOR ANY-SCENARIO PERSON RE-IDENTIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

In Any-Scenario Person Re-Identification (AS-ReID), a unified model must retrieve identities across spectral modalities and aerial–ground views. However, existing methods handle these variations through joint scenario alignment without explicitly accounting for their distinct characteristics, potentially limiting retrieval performance. To address this limitation, we propose OSFAL, which combines modality-specific orthogonal alignment with layer-localized cross-view learning to mitigate spectral and view discrepancies through distinct mechanisms. Specifically, Orthogonal Cross-Spectral Feature Alignment (OCSFA) combines orthogonal rotations with mean alignment to map features from different spectral modalities into a shared reference coordinate system while preserving within-modality Euclidean distances. We learn these transformations by encouraging the modality–view prototypes of each identity to agree with a identity consensus prototype. Layer-Localized Cross-View Learning (LLCVL) compares within-view and cross-view similarities across encoder blocks, together with cross-view identity matching accuracy, to locate a key supervision boundary. Bidirectional cross-view contrastive learning is then applied at the selected block output within each modality. Experiments on WHU-MARS show that OSFAL improves retrieval performance across different evaluation protocols, achieving +5.0% gain in Rank-1 accuracy over the baseline on WHU-MARS-1000.

1 INTRODUCTION

Person re-identification (ReID) aims to retrieve observations of the same person across non-overlapping cameras (Ye et al., 2021; Luo et al., 2019; He et al., 2024; Leng et al., 2019; Lin et al., 2025). Cross-modality ReID aims to match identities across different spectral modalities by reducing sensor-dependent appearance discrepancies (Zuo et al., 2026; Zheng et al., 2022; Kim et al., 2023; Wang et al., 2022; Yang et al., 2024; Jiang et al., 2022; Ren & Zhang, 2024). Aerial–ground ReID focuses on matching identities between ground cameras and aerial platforms despite changes in viewpoint, scale, and occlusion (Zhang et al., 2024; Wang et al., 2025; Nguyen et al., 2024; Wang et al., 2026; Zheng et al., 2026; Li et al., 2026). In Any-Scenario Person Re-Identification (AS-ReID), these challenges coexist: a single model must match a query from any spectral modality and viewpoint against a heterogeneous gallery containing all available conditions (Zhao et al., 2026b). The representation must therefore accommodate substantial appearance changes within each identity while distinguishing different identities across scenarios.

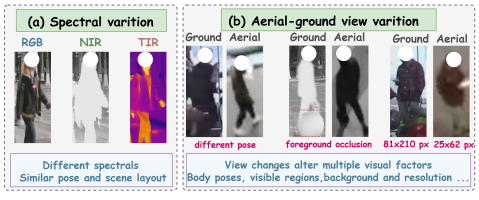

Figure 1: Examples of spectral and view variations. (a) Paired spectral observations share similar pose and scene layout. (b) Aerial–ground observations of the same identity differ in pose, occlusion, and image resolution.

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

However, existing scenario-level alignment methods primarily mitigate spectral and view discrepancies through feature aggregation and identity-level alignment (Zhao et al., 2026b). This design lacks alignment mechanisms tailored to the different structures of spectral and view variations, potentially limiting retrieval performance (Ha et al., 2025; Wang et al., 2024). As illustrated in Fig. 1, spectral and view variations arise from distinct imaging mechanisms. Paired spectral observations largely preserve viewpoint, pose, and scene content while changing the sensing spectrum (Wu et al., 2017; Nguyen et al., 2017; Zhang & Wang, 2023), whereas Ground and Aerial observations introduce spatial misalignment, scale differences, and varying occlusions (Khalid et al., 2025; Yang et al., 2026). This distinction motivates alignment mechanisms tailored to each source of variation. We therefore propose Orthogonal cross-Spectral Feature Alignment and Layer-localized cross-view learning (OSFAL) method, which combines modality-specific orthogonal alignment with layer-localized cross-view learning to mitigate spectral and view discrepancies through distinct mechanisms.

We introduce Orthogonal Cross-Spectral Feature Alignment (OCSFA), motivated by AS-ReID evaluation using cosine similarity or Euclidean distance between ℓ_2-normalized features. Since test-time retrieval depends on feature directions rather than magnitudes, we focus on adjusting feature orientations. Orthogonal rotations provide a natural mechanism: by their algebraic properties, they change vector directions without altering vector lengths (Trockman & Kolter, 2021; Qiu et al., 2023). We therefore learn modality-specific orthogonal rotations to adjust cross-spectral relationships while preserving norms and within-modality angles. To also align feature centers, OCSFA subtracts the source-modality mean, rotates the centered features into the reference coordinates, and adds the reference mean. This affine transformation preserves within-modality Euclidean distances before normalization. We learn the transformations by encouraging the modality–view prototypes of each identity to agree with an identity consensus prototype (Shi et al., 2024a).

A natural extension of this idea is to introduce view-specific orthogonal rotations. However, aerial–ground differences involve changes in body orientation, occlusion, spatial layout, and image resolution, making them difficult to model using a single global orthogonal transformation. Related methods tackle this complexity through view-decoupled architectures and adaptive prompts (Zhang et al., 2024; Wang et al., 2025), highlighting the importance of learning view-robust representations within the network. We therefore introduce Layer-Localized Cross-View Learning (LLCVL). Specifically, we construct block-wise statistics to examine how representations evolve along the encoder depth (Raghu et al., 2021). Our analysis suggests that within-view identity consistency develops earlier, whereas substantial reductions in cross-view discrepancy occur in deeper blocks (Raghu et al., 2021). Guided by this observation, LLCVL locates the key block boundary for cross-view learning. We then apply bidirectional cross-view contrastive supervision at this boundary within each modality to improve aerial–ground identity matching (Radford et al., 2021; Shi et al., 2024b; Yu et al., 2024). Together, OCSFA and LLCVL combine spectral alignment at the encoder output with targeted cross-view supervision within the network.

Our main contributions are summarized as follows: (1) We propose OSFAL, a unified AS-ReID framework that explicitly handles spectral and view discrepancies through distinct, complementary mechanisms. (2) We introduce OCSFA, which combines modality-specific orthogonal rotations, mean alignment, and identity-wise prototype consistency while preserving within-modality Euclidean distances before normalization. (3) We develop LLCVL, which locates a key block boundary in the network and applies bidirectional cross-view contrastive supervision within each modality, without adding an inference branch. (4) Experiments on WHU-MARS demonstrate consistent improvements across multiple evaluation protocols, including an approximately 5.0 % Rank-1 gain over the UAD baseline on WHU-MARS-1000.

An extended discussion of related work is provided in Appendix A.

2 METHOD

OSFAL models spectral and view variations through two distinct mechanisms. Orthogonal Cross-Spectral Feature Alignment (OCSFA) maps modality-specific features into a common spectral coordinate system while preserving their within-modality geometry. Layer-Localized Cross-View Learning (LLCVL) identifies the network boundary at which view-dependent representations begin to change and then applies cross-view supervision at that boundary. Figure 2 illustrates the framework.

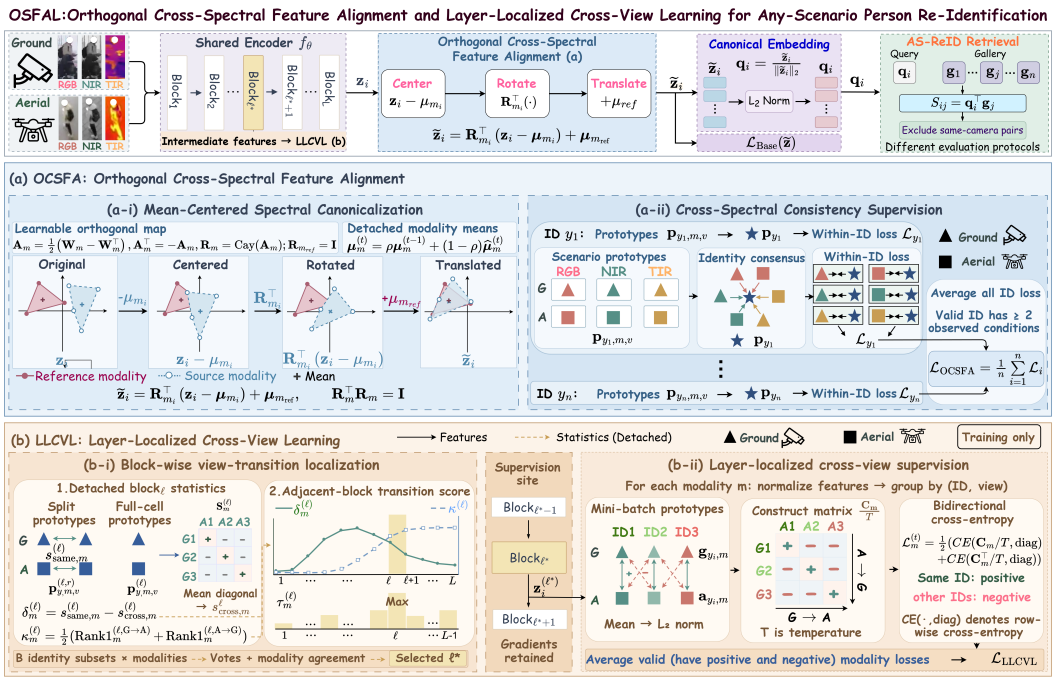

Figure 2: Overview of the proposed OSFAL framework. (a) OCSFA aligns spectral features through mean centering, orthogonal rotation, and translation to the reference mean, and enforces prototype consistency within each identity. (b) LLCVL uses detached statistics to select a block boundary by voting, then applies bidirectional cross-view contrastive supervision within each modality. Inference uses the normalized OCSFA output; LLCVL is training-only.

2.1 PRELIMINARIES

2.1.1 AS-REID FORMULATION

Let $\mathcal{M} = \{\mathrm{RGB}, \mathrm{NIR}, \mathrm{TIR}\}$ and $\mathcal{V} = \{\mathrm{G}, \mathrm{A}\}$ denote the modality and view sets, respectively. Their Cartesian product $\mathcal{S} = \mathcal{M} \times \mathcal{V}$ defines the scenario space. Each sample is represented as $(\mathbf{x}_i, y_i, s_i, c_i)$, where $\mathbf{x}_i \in \mathcal{X}$ is an image, $y_i \in \mathcal{Y}$ is its identity, $s_i = (m_i, v_i) \in \mathcal{S}$ is its scenario with $m_i \in \mathcal{M}$ and $v_i \in \mathcal{V}$, and $c_i \in \mathcal{K}$ is its camera index. A shared encoder $f_\theta : \mathcal{X} \to \mathbb{R}^d$ produces the retrieval feature $\mathbf{z}_i = f_\theta(\mathbf{x}_i)$. We instantiate f_θ with either a ViT (Dosovitskiy et al., 2020), using its final class token, or a ResNet (He et al., 2016), using its global-average-pooled output. Retrieval uses cosine similarity or, equivalently for ℓ_2-normalized features, Euclidean distance.

Let $\mathcal{Q}$ and $\mathcal{G}$ denote the query and gallery sets, respectively. Under the AS-ReID protocol, both sets span the complete scenario space $\mathcal{S}$, and each query is matched against samples from all admissible scenarios after excluding those captured by the same camera. Conventional protocols, including VI-ReID, AG-ReID, MM-ReID, and traditional ReID, are obtained by restricting the scenario subsets of $\mathcal{Q}$ and $\mathcal{G}$ accordingly.

2.1.2 ALGEBRAIC FOUNDATIONS OF ORTHOGONAL TRANSFORMATIONS

Orthogonal isometries. A matrix $\mathbf{R} \in \mathbb{R}^{d\times d}$ is orthogonal if $\mathbf{R}^\top\mathbf{R} = \mathbf{I}_d$. For any two vectors $\mathbf{a}_i, \mathbf{a}_j \in \mathbb{R}^d$, applying the same orthogonal transformation preserves their inner product and Euclidean distance:

$$(\mathbf{R}\mathbf{a}_i)^\top(\mathbf{R}\mathbf{a}_j) = \mathbf{a}_i^\top\mathbf{a}_j, \qquad \|\mathbf{R}\mathbf{a}_i - \mathbf{R}\mathbf{a}_j\|_2 = \|\mathbf{a}_i - \mathbf{a}_j\|_2 . \tag{1}$$

Orthogonal transformations also preserve vector norms and cosine similarities. They therefore act as isometries, reorienting the coordinate system without scaling, shearing, or collapsing the original feature geometry.

Dense Cayley parameterization. The Cayley transform provides a differentiable parameterization of orthogonal matrices (Trockman & Kolter, 2021; Qiu et al., 2025). Given an unconstrained matrix $\mathbf{W} \in \mathbb{R}^{d\times d}$, we first construct a skew-symmetric generator

$$\mathbf{A} = \frac{1}{2}\left(\mathbf{W} - \mathbf{W}^\top\right), \qquad \mathbf{A}^\top = -\mathbf{A}. \tag{2}$$

Its Cayley transform is defined as $\mathbf{R} = \mathrm{Cay}(\mathbf{A}) = (\mathbf{I}_d + \mathbf{A})^{-1}(\mathbf{I}_d - \mathbf{A})$. For a real skew-symmetric $\mathbf{A}$, $\mathbf{I}_d + \mathbf{A}$ is invertible, and the resulting matrix satisfies

$$\mathbf{R}^\top\mathbf{R} = \mathbf{I}_d, \qquad \mathbf{R}^{-1} = \mathbf{R}^\top. \tag{3}$$

Consequently, $\mathrm{rank}(\mathbf{R}) = d$. Detailed derivations of the isometric properties and the dense Cayley parameterization are provided in Appendix B.

2.2 Orthogonal Cross-Spectral Feature Alignment

OCSFA maps features from different spectral modalities into a common coordinate system. Specifically, it combines a modality-specific orthogonal transformation with mean translation: the former corrects differences in feature orientation while preserving the geometry established in Equation 1, and the latter aligns modality-dependent feature centers. We first define the orthogonal mapping with respect to a reference modality and then derive the complete canonicalization operation.

2.2.1 Reference-modality orthogonal mapping

Among the three spectral modalities, we select an arbitrary modality $m_{\mathrm{ref}} \in \mathcal{M}$ as the reference and use its feature coordinate system as the canonical spectral space. Each modality $m \in \mathcal{M}$ is associated with an orthogonal matrix $\mathbf{R}_m \in \mathbb{R}^{d\times d}$, and the reference transformation is fixed as $\mathbf{R}_{m_{\mathrm{ref}}} = \mathbf{I}_d$. For every non-reference modality, $\mathbf{R}_m$ is constructed from a learnable skew-symmetric generator using the Cayley transform in Equation 2.1.2. We define the direction of $\mathbf{R}_m$ from the reference coordinate system to that of modality m. Therefore, a feature direction expressed in modality-m coordinates is mapped back to the canonical system by $\mathbf{R}_m^{-1}$. Orthogonality gives $\mathbf{R}_m^{-1} = \mathbf{R}_m^\top$, avoiding the need to learn a separate inverse transformation.

2.2.2 Mean-centered spectral canonicalization

The orthogonal matrices model differences in feature orientation. However, rotating a feature directly about the origin also rotates the center of its modality distribution. In general, the resulting center $\mathbf{R}_m^\top\boldsymbol{\mu}_m$ does not coincide with the center of the reference modality. Cross-spectral canonicalization must therefore account for both the orientation and center of each modality distribution.

At iteration t, let $\mathcal{B}^{(t)}$ denote the set of sample indices in the current mini-batch and define its modality-m subset as $\mathcal{B}_m^{(t)} = \left\{i \in \mathcal{B}^{(t)} \mid m_i = m\right\}$. The corresponding batch mean is $\widehat{\boldsymbol{\mu}}_m^{(t)} = \frac{1}{|\mathcal{B}_m^{(t)}|}\sum_{i\in\mathcal{B}_m^{(t)}} \mathrm{sg}[\mathbf{z}_i]$, where $\mathrm{sg}[\cdot]$ denotes stop-gradient. We update the running feature center of modality m using $\boldsymbol{\mu}_m^{(t)} = \rho\boldsymbol{\mu}_m^{(t-1)} + (1-\rho)\widehat{\boldsymbol{\mu}}_m^{(t)}$, where $\rho \in [0, 1)$ is the momentum coefficient. A modality mean is initialized by its first available batch estimate and remains unchanged when $\mathcal{B}_m^{(t)}$ is empty.

Given these modality statistics, we construct the canonicalization in three conceptual steps. We first subtract the source-modality center so that the orthogonal transformation acts on the sample displacement rather than on its absolute position. We then rotate this centered displacement into the reference coordinate system using $\mathbf{R}_{m_i}^\top$. Finally, we add the reference-modality center, yielding

$$\widetilde{\mathbf{z}}_i^{(t)} = \mathbf{R}_{m_i}^\top\left(\mathbf{z}_i - \boldsymbol{\mu}_{m_i}^{(t)}\right) + \boldsymbol{\mu}_{m_{\mathrm{ref}}}^{(t)}. \tag{4}$$

This construction aligns the modality centers by design. Treating the running means as estimates of their population counterparts gives

$$\mathbb{E}\left[\widetilde{\mathbf{z}}_i^{(t)} \mid m_i = m\right] = \mathbf{R}_m^\top\left(\mathbb{E}[\mathbf{z}_i \mid m_i = m] - \boldsymbol{\mu}_m^{(t)}\right) + \boldsymbol{\mu}_{m_{\mathrm{ref}}}^{(t)} \approx \boldsymbol{\mu}_{m_{\mathrm{ref}}}^{(t)}. \tag{5}$$

For two samples i and j from the same modality m, the translation terms cancel, and orthogonality gives

$$\widetilde{\mathbf{z}}_i^{(t)} - \widetilde{\mathbf{z}}_j^{(t)} = \mathbf{R}_m^\top(\mathbf{z}_i - \mathbf{z}_j), \qquad \left\|\widetilde{\mathbf{z}}_i^{(t)} - \widetilde{\mathbf{z}}_j^{(t)}\right\|_2 = \|\mathbf{z}_i - \mathbf{z}_j\|_2 . \tag{6}$$

Equation 5 shows that the canonicalized features of different modalities share the reference-modality mean, whereas Equation 6 shows that pairwise distances within each modality remain unchanged. Together, these properties allow the proposed affine transformation to align first-order modality statistics without distorting the within-modality metric structure.

2.2.3 Cross-spectral consistency supervision

The orthogonality constraint restricts each modality transformation to preserve feature geometry, but it does not determine which orthogonal matrix establishes identity-consistent correspondences across spectral modalities. Indeed, multiple orthogonal matrices can preserve the same within-modality relationships while inducing substantially different cross-spectral similarities. We therefore learn the modality-specific transformations by encouraging canonicalized representations of the same identity to remain consistent across the observed modality–view conditions.

Because AS-ReID retrieval is based on angular similarity, we first ℓ_2-normalize the canonicalized feature: $\mathbf{q}_i^{(t)} = \mathrm{norm}\left(\widetilde{\mathbf{z}}_i^{(t)}\right)$, where $\mathrm{norm}(\mathbf{a}) = \mathbf{a}/\|\mathbf{a}\|_2$ and $\mathbf{q}_i^{(t)} \in \mathbb{R}^d$ is the normalized canonicalized representation of sample i. Directly minimizing distances between individual cross-spectral samples can be sensitive to pose, background, and instance-level noise (Wang & Isola, 2020). We instead aggregate samples according to their identity, modality, and view before applying the consistency constraint (Wang et al., 2021). For identity y, modality m, and view v, let

$$\mathcal{B}_{y,m,v}^{(t)} = \left\{ i \in \mathcal{B}^{(t)} \;\middle|\; y_i = y,\, m_i = m,\, v_i = v \right\} \tag{7}$$

denote the corresponding subset of the current mini-batch. For every nonempty subset, we first compute its mean canonicalized representation and then normalize it to obtain the cell prototype (Snell et al., 2017):

$$\overline{\mathbf{q}}_{y,m,v}^{(t)} = \frac{1}{\left|\mathcal{B}_{y,m,v}^{(t)}\right|} \sum_{i \in \mathcal{B}_{y,m,v}^{(t)}} \mathbf{q}_i^{(t)}, \qquad \mathbf{p}_{y,m,v}^{(t)} = \frac{\overline{\mathbf{q}}_{y,m,v}^{(t)}}{\left\|\overline{\mathbf{q}}_{y,m,v}^{(t)}\right\|_2}. \tag{8}$$

Here, $\mathbf{p}_{y,m,v}^{(t)} \in \mathbb{R}^d$ summarizes identity y under the modality–view condition (m, v). The view index stratifies the prototypes, whereas $\mathbf{R}_m$ remains shared across views.

Let $\mathcal{D}_y^{(t)} = \left\{ (m,v) \in \mathcal{M} \times \mathcal{V} \;\middle|\; \left|\mathcal{B}_{y,m,v}^{(t)}\right| > 0 \right\}$ denote the set of modality–view conditions observed for identity y in the current mini-batch. We aggregate their prototypes into an identity-level consensus representation:

$$\mathbf{p}_y^{(t)} = \mathrm{norm}\,(\frac{1}{\left|\mathcal{D}_y^{(t)}\right|} \sum_{(m,v) \in \mathcal{D}_y^{(t)}} \mathbf{p}_{y,m,v}^{(t)}), \tag{9}$$

where $\mathrm{norm}(\mathbf{a}) = \mathbf{a}/\|\mathbf{a}\|_2$. The consensus prototype provides a common identity anchor constructed from all available conditions rather than privileging a particular non-reference modality.

Only identities observed under at least two conditions provide a valid consistency signal. We therefore define $\mathcal{Y}_+^{(t)} = \left\{ y \;\middle|\; \left|\mathcal{D}_y^{(t)}\right| \geq 2 \right\}$. The cross-spectral consistency objective minimizes the cosine distance between each condition-specific prototype and its identity consensus:

$$\mathcal{L}_{\mathrm{OCSFA}}^{(t)} = \frac{1}{|\mathcal{Y}_+^{(t)}|} \sum_{y \in \mathcal{Y}_+^{(t)}} \frac{1}{|\mathcal{D}_y^{(t)}|} \sum_{(m,v) \in \mathcal{D}_y^{(t)}} \left[1 - \left(\mathbf{p}_{y,m,v}^{(t)}\right)^\top \mathbf{p}_y^{(t)} \right]. \tag{10}$$

The loss is computed on prototypes obtained by aggregating samples with the same identity, modality, and view, rather than on individual image pairs. This aggregation reduces sensitivity to image-specific variations, such as imperfect cropping, local occlusion, and background changes, and allows the consistency objective to be applied without explicitly constructing synchronized cross-spectral pairs (Liu et al., 2020). Combined with the orthogonality constraint, the objective learns modality transformations that improve identity consistency after spectral canonicalization while preserving the within-modality feature geometry.

2.3 LAYER-LOCALIZED CROSS-VIEW LEARNING

Unlike spectral variation, the discrepancy between Ground and Aerial views involves spatial misalignment, scale changes, occlusion, background, and resolution differences. Such compound variation may be modeled differently across network depths, making the choice of supervision layer important (Raghu et al., 2021). LLCVL therefore locates the block boundary at which cross-view representations undergo the most significant transition and applies cross-view supervision at this boundary. Let $\mathbf{z}_i^{(\ell)} \in \mathbb{R}^{d_\ell}$ denote the representation of sample i after block $\ell \in \{1, \dots, L\}$. The localization statistics are computed without gradient propagation, whereas the subsequent cross-view objective directly supervises the selected intermediate representation.

2.3.1 BLOCK-WISE VIEW-TRANSITION LOCALIZATION

For each block and spectral modality, we characterize the representation by within-view representation stability and cross-view identity consistency. All statistics are computed from detached, normalized features $\overline{\mathbf{z}}_i^{(\ell)} = \mathrm{norm}(\mathrm{sg}[\mathbf{z}_i^{(\ell)}])$, $\mathrm{sg}[\cdot]$ denotes stop-gradient.

Let $\mathcal{C}_{y,m,v}$ contain the observations accumulated during an epoch for identity y, modality m, and view v. Their block-level prototype is $\mathbf{p}_{y,m,v}^{(\ell)} = \mathrm{norm}\,(\frac{1}{|\mathcal{C}_{y,m,v}|}\sum_{i\in\mathcal{C}_{y,m,v}} \overline{\mathbf{z}}_i^{(\ell)})$. To independently estimate within-view stability, we alternately divide each cell into two disjoint subsets, $\mathcal{C}_{y,m,v} = \mathcal{C}_{y,m,v}^{(1)} \dot\cup \mathcal{C}_{y,m,v}^{(2)}$, and construct $\mathbf{p}_{y,m,v}^{(\ell,r)} = \mathrm{norm}\,(\frac{1}{|\mathcal{C}_{y,m,v}^{(r)}|}\sum_{i\in\mathcal{C}_{y,m,v}^{(r)}} \overline{\mathbf{z}}_i^{(\ell)}), r \in \{1,2\}$. The full prototype is used for cross-view comparison, whereas the split prototypes are used only to estimate within-view stability.

Let $\mathcal{Y}_m$ contain identities with valid split prototypes in both Ground and Aerial views under modality m. We construct the cross-view similarity matrix

$$\mathbf{S}_m^{(\ell)}(y, y') = \big(\mathbf{p}_{y,m,\mathrm{G}}^{(\ell)}\big)^\top \mathbf{p}_{y',m,\mathrm{A}}^{(\ell)}, \qquad y, y' \in \mathcal{Y}_m. \tag{11}$$

Its diagonal contains same-identity cross-view similarities, while its off-diagonal elements compare different identities. We measure cross-view identity reliability by the bidirectional rank-1 accuracy

$$\kappa_m^{(\ell)} = \frac{1}{2|\mathcal{Y}_m|}\sum_{y\in\mathcal{Y}_m}\Big[\mathbb{I}\,(\mathbf{S}_m^{(\ell)}(y,y) > \max_{y'\neq y}\mathbf{S}_m^{(\ell)}(y,y')) + \mathbb{I}\,(\mathbf{S}_m^{(\ell)}(y,y) > \max_{y'\neq y}\mathbf{S}_m^{(\ell)}(y',y))\Big]. \tag{12}$$

The two indicators evaluate Ground-to-Aerial and Aerial-to-Ground retrieval, respectively.

We quantify the remaining view discrepancy by comparing within-view stability with same-identity cross-view similarity:

$$\begin{aligned} s_{\mathrm{same},m}^{(\ell)} &= \frac{1}{2|\mathcal{Y}_m|}\sum_{y\in\mathcal{Y}_m}\sum_{v\in\{\mathrm{G,A}\}}\big(\mathbf{p}_{y,m,v}^{(\ell,1)}\big)^\top \mathbf{p}_{y,m,v}^{(\ell,2)}, \\ s_{\mathrm{cross},m}^{(\ell)} &= \frac{1}{|\mathcal{Y}_m|}\sum_{y\in\mathcal{Y}_m}\mathbf{S}_m^{(\ell)}(y,y), \qquad \delta_m^{(\ell)} = s_{\mathrm{same},m}^{(\ell)} - s_{\mathrm{cross},m}^{(\ell)}. \end{aligned} \tag{13}$$

A positive $\delta_m^{(\ell)}$ indicates that same-identity representations are more consistent within a view than across views. The transition score between blocks ℓ and $\ell+1$ is

$$\tau_m^{(\ell)} = \sqrt{\kappa_m^{(\ell)}\kappa_m^{(\ell+1)}}\begin{cases}\mathbb{I}[\delta_m^{(\ell)} > 0]\,\mathbb{I}[\delta_m^{(\ell+1)} \le 0], & \text{if a sign crossing exists,}\\ [\delta_m^{(\ell)} - \delta_m^{(\ell+1)}]_+, & \text{otherwise,}\end{cases} \tag{14}$$

where $[a]_+ = \max(a, 0)$. Here, $\tau_m^{(\ell)}$ evaluates the transition from block ℓ to block $\ell+1$: a large value indicates that the transition is identity-reliable and that the residual view gap either crosses zero or decreases sharply.

To reduce sensitivity to identity composition, we compute these statistics over B fixed identity subsets. For each subset b, we aggregate $\kappa_{b,m}^{(\ell)}$ and $\delta_{b,m}^{(\ell)}$ across modalities to obtain $\overline{\tau}_b^{(\ell)}$. Each subset nominates one boundary, and the final boundary is selected by voting:

$$\ell_b^\star = \operatorname*{arg\,max}_{\ell\in\{1,\dots,L-1\}}\overline{\tau}_b^{(\ell)}, \qquad \ell^\star = \operatorname*{arg\,max}_{\ell\in\{1,\dots,L-1\}}\sum_{b=1}^{B}\mathbb{I}(\ell_b^\star = \ell). \tag{15}$$

We additionally require modality-specific agreement before applying LLCVL to $\mathbf{z}_i^{(\ell^\star)}$, the output of block $\ell^\star$. Partitioning, consensus, and activation details are provided in Appendix C.

2.3.2 LAYER-LOCALIZED CROSS-VIEW SUPERVISION

The selected boundary $\ell^\star$ lies between blocks $\ell^\star$ and $\ell^\star + 1$, and cross-view supervision is applied to the output of block $\ell^\star$. Unlike the detached features used for localization, the training features retain their gradients. We define $\mathbf{u}_i^{(t)} = \mathrm{norm}(\mathbf{z}_i^{(\ell^\star)})$ and let $\mathcal{B}_{y,m,v}^{(t)}$ contain the samples of identity y, modality m, and view v in the current mini-batch. The Ground and Aerial prototypes are

$$\mathbf{g}_{y,m}^{(t)} = \mathrm{norm}\,(\frac{1}{|\mathcal{B}_{y,m,\mathrm{G}}^{(t)}|}\sum_{i\in\mathcal{B}_{y,m,\mathrm{G}}^{(t)}}\mathbf{u}_i^{(t)}), \qquad \mathbf{a}_{y,m}^{(t)} = \mathrm{norm}\,(\frac{1}{|\mathcal{B}_{y,m,\mathrm{A}}^{(t)}|}\sum_{i\in\mathcal{B}_{y,m,\mathrm{A}}^{(t)}}\mathbf{u}_i^{(t)}). \tag{16}$$

Prototypes are constructed separately within each modality so that the objective focuses on view variation without mixing spectral discrepancies.

Let $\mathcal{Y}_m^{(t)}$ contain identities for which both prototypes are available. Using the same identity ordering for both views, we construct $\mathbf{C}_m^{(t)}(y,y') = \big(\mathbf{g}_{y,m}^{(t)}\big)^\top \mathbf{a}_{y',m}^{(t)}, y, y' \in \mathcal{Y}_m^{(t)}$. The diagonal entries compare the same identity across views, whereas the off-diagonal entries provide cross-identity negatives (Khosla et al., 2020). We optimize both retrieval directions using

$$\mathcal{L}_m^{(t)} = \frac{1}{2}\left[\mathrm{CE}\Big(\mathbf{C}_m^{(t)}/T, \mathrm{diag}\Big) + \mathrm{CE}\Big((\mathbf{C}_m^{(t)})^\top/T, \mathrm{diag}\Big)\right], \tag{17}$$

where T is the temperature and $\mathrm{CE}(\cdot, \mathrm{diag})$ denotes row-wise cross-entropy using the diagonal entry as the positive class. The first term retrieves Aerial prototypes using Ground queries, while the second performs the reverse retrieval. An explicit expansion of the $\mathcal{L}_m^{(t)}$ is provided in Appendix D.

Only modalities containing at least two valid identities provide both positive and negative pairs. Defining $\mathcal{M}_{\mathrm{act}}^{(t)} = \{m \in \mathcal{M} \mid |\mathcal{Y}_m^{(t)}| \geq 2\}$, the final objective is

$$\mathcal{L}_{\mathrm{LLCVL}}^{(t)} = \frac{1}{|\mathcal{M}_{\mathrm{act}}^{(t)}|}\sum_{m\in\mathcal{M}_{\mathrm{act}}^{(t)}}\mathcal{L}_m^{(t)}. \tag{18}$$

The loss is set to zero if no modality is active.

2.4 OVERALL TRAINING OBJECTIVE

The complete objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{Base}}(\widetilde{\mathbf{z}}) + \lambda_{\mathrm{O}}\mathcal{L}_{\mathrm{OCSFA}} + \lambda_{\mathrm{L}}\mathcal{L}_{\mathrm{LLCVL}}, \tag{19}$$

where $\mathcal{L}_{\mathrm{Base}}$ is the baseline ReID objective, and λ_{O} and λ_{L} weight the spectral and view losses. LLCVL is used only during training. At inference, retrieval uses the final OCSFA-canonicalized embedding and introduces no additional branch.

3 EXPERIMENTS

3.1 EXPERIMENTAL SETUP

Datasets and evaluation metrics. We conduct experiments on WHU-MARS (Zhao et al., 2026b), which, to our knowledge, is currently the only benchmark dedicated to AS-ReID. It provides two versions, WHU-MARS-1000 and WHU-MARS-2337, for which we follow the official splits and evaluation protocols. We report CMC accuracies at Rank-1, Rank-5, and Rank-10, along with mean average precision (mAP), as evaluation metrics.

Implementation details. We adopt UAD, introduced alongside the WHU-MARS benchmark in the same work (Zhao et al., 2026b), as our baseline with a ViT-B/16 backbone. RGB serves as the reference modality, with $\mathbf{R}_{\mathrm{RGB}} = \mathbf{I}_d$. We set the EMA momentum for modality means to $\rho = 0.99$ and the LLCVL contrastive temperature to $T = 0.2$. The OCSFA and LLCVL loss weights are set to $\lambda_{\mathrm{O}} = 0.1$ and $\lambda_{\mathrm{L}} = 0.05$, respectively. All other backbone, data preprocessing, optimization, and evaluation settings follow the baseline. All experiments are conducted on four NVIDIA GeForce RTX 4080 SUPER GPUs.

Table 1: Comparison under the AS-ReID setting. WHU-MARS-1000 and WHU-MARS-2337 are trained and evaluated on their respective splits. WHU-MARS-1000-GD evaluates the WHU-MARS-1000-trained model using ground and daytime queries and galleries only. **Bold** values indicate the best performance. * denotes our reproduction.

Method	Venue	Params	WHU-MARS-1000-GD				WHU-MARS-1000				WHU-MARS-2337			
			mAP	R-1	R-5	R-10	mAP	R-1	R-5	R-10	mAP	R-1	R-5	R-10
BoT (Luo et al., 2019)	CVPRW19	23.5M	10.6	24.9	37.9	45.6	5.9	18.4	32.5	40.1	4.6	14.8	26.2	32.5
AGW (Ye et al., 2021)	TPAMI21	23.5M	6.1	17.0	27.8	35.9	4.0	15.6	28.7	36.0	5.2	16.9	28.7	35.1
MSINet (Gu et al., 2023)	CVPR23	2.3M	9.7	26.9	39.5	46.3	5.6	21.1	37.5	46.0	2.6	14.8	25.4	31.6
TransReID (He et al., 2021)	ICCV21	99.9M	9.6	18.9	34.1	43.4	7.4	17.1	30.3	38.2	5.5	13.1	23.4	29.1
TransReID-SSL (Luo et al., 2021)	ArXiv21	88.4M	14.0	30.1	44.1	51.7	10.1	26.7	43.8	51.7	9.2	24.4	39.1	46.2
DC-Former (Li et al., 2023b)	AAAI23	85.8M	9.9	20.1	34.9	42.0	6.5	16.3	29.9	36.7	4.3	10.9	19.8	25.5
CLIP-ReID (Li et al., 2023a)	AAAI23	125.3M	13.8	28.9	42.7	49.6	10.6	26.6	42.6	50.4	9.3	24.1	37.5	44.0
PHA (Zhang et al., 2023)	CVPR23	107.0M	9.6	22.0	33.9	42.3	7.5	15.9	28.2	35.3	5.5	12.0	21.1	26.3
VDT (Zhang et al., 2024)	CVPR24	85.8M	13.1	31.0	44.1	51.1	7.8	25.2	41.9	50.4	6.3	20.0	33.5	40.4
DTST (Wang & Pishgar, 2024)	ArXiv24	86.9M	11.6	29.8	42.5	47.9	6.8	23.4	39.8	47.9	7.5	23.5	37.9	45.3
SeCap (Wang et al., 2025)	CVPR25	130.9M	13.4	28.1	41.2	47.8	10.4	26.4	42.4	50.4	8.0	21.4	34.1	41.0
UAD* (Zhao et al., 2026b)	CVPR26	85.7M	14.4	31.8	47.1	52.6	11.0	29.3	47.1	54.6	9.5	25.9	40.9	48.2
OSFAL (Ours)	This paper	88.7M	**16.0**	**35.9**	**50.7**	**57.9**	**12.3**	**34.3**	**51.4**	**58.8**	**10.4**	**28.8**	**43.5**	**50.8**

Table 2: AG-ReID comparison on WHU-MARS-1000. A/G denote Aerial/Ground, respectively. **Bold** values indicate the best performance. * denotes our reproduction.

Method	A → G			G → A		
	mAP	R-1	R-10	mAP	R-1	R-10
VDT	8.9	30.4	56.7	10.7	30.2	56.4
DTST	7.9	30.5	56.2	9.8	28.5	57.3
SeCap	10.1	26.4	50.4	12.0	25.3	50.7
UAD* (Baseline)	11.4	32.5	58.3	13.1	30.3	55.1
OSFAL (Ours)	**13.5**	**39.4**	**63.6**	**15.3**	**36.0**	**60.7**

Table 3: Bidirectional VI-ReID comparison on WHU-MARS-1000. UAD denotes our reproduced baseline. **Bold** values indicate the best performance. * denotes our reproduction.

Query → Gallery	UAD* (Baseline)			OSFAL (Ours)		
	R-1	mAP	mINP	R-1	mAP	mINP
RGB → NIR	26.23	15.96	2.28	**27.17**	**17.33**	**2.51**
NIR → RGB	24.64	15.62	2.85	**25.57**	**16.96**	**3.16**
RGB → TIR	6.89	5.59	1.12	**7.63**	**5.88**	**1.13**
TIR → RGB	7.73	**5.89**	1.31	**8.52**	5.86	**1.35**

3.2 Main results

Comparison with state-of-the-art methods. (1) **AS-ReID.** Table 1 compares OSFAL with existing ReID methods under the AS-ReID setting. OSFAL achieves the best performance among the compared methods across all four metrics in the three evaluated settings. On WHU-MARS-1000, it reaches 12.3% mAP and 34.3% Rank-1, outperforming UAD by 1.3 and 5.0 percentage points, respectively. On WHU-MARS-1000-GD and WHU-MARS-2337, OSFAL achieves 35.9% and 28.8% Rank-1, improving over UAD by 4.1 and 2.9 percentage points, respectively. These results demonstrate consistent retrieval improvements under both full-scenario and Ground-daytime evaluation. (2) **AG-ReID.** Table 2 reports aerial–ground retrieval results on WHU-MARS-1000. OSFAL achieves the best mAP, Rank-1, and Rank-10 among the compared methods in both retrieval directions. Compared with the reproduced UAD baseline, Rank-1 increases from 32.5% to 39.4% for Aerial-to-Ground retrieval and from 30.3% to 36.0% for Ground-to-Aerial retrieval. The improvements in both mAP and rank-based metrics indicate that OSFAL benefits aerial–ground matching beyond the full-scenario evaluation. (3) **VI-ReID.** Table 3 compares OSFAL with the reproduced UAD baseline in four visible–infrared retrieval directions. Both methods use their AS-ReID-trained checkpoints without protocol-specific retraining. OSFAL improves Rank-1 and mINP in all four directions, while increasing mAP in three. For RGB-to-NIR and NIR-to-RGB retrieval, mAP rises from 15.96% and 15.62% to 17.33% and 16.96%, respectively. For RGB-to-TIR and TIR-to-RGB retrieval, Rank-1 increases from 6.89% and 7.73% to 7.63% and 8.52%, respectively.

Ablation study. Table 4 evaluates the individual and combined contributions of OCSFA and LLCVL. OCSFA independently improves mAP and Rank-1 in all three settings, whereas LLCVL mainly benefits protocols involving both Ground and Aerial views. Its limited effect on WHU-MARS-1000-GD is expected, as this protocol contains only Ground images and involves no aerial–ground matching. On WHU-MARS-1000, adding OCSFA increases Rank-1 from 29.3% to 33.4%,

Table 4: Component ablation of OSFAL under the AS-ReID setting. OCSFA and LLCVL denote orthogonal cross-spectral feature alignment and layer-localized cross-view learning, respectively. **Bold** values indicate the best performance.

Variant	Components		WHU-MARS-1000-GD				WHU-MARS-1000				WHU-MARS-2337			
	OCSFA	LLCVL	mAP	R-1	R-5	R-10	mAP	R-1	R-5	R-10	mAP	R-1	R-5	R-10
UAD (Baseline)	✗	✗	14.4	31.8	47.0	52.6	11.0	29.3	47.1	54.6	9.5	25.9	40.9	48.2
UAD + OCSFA	✓	✗	15.8	35.7	50.0	57.5	12.0	33.4	50.3	58.6	10.1	28.0	42.7	50.3
UAD + LLCVL	✗	✓	14.5	32.2	47.4	53.8	11.2	30.0	46.8	55.1	9.8	26.5	41.6	49.1
UAD + OSFAL	✓	✓	**16.0**	**35.9**	**50.7**	**57.9**	**12.3**	**34.3**	**51.4**	**58.8**	**10.4**	**28.8**	**43.5**	**50.8**

while LLCVL alone raises it to 30.0%. Combining both modules further improves Rank-1 to 34.3%, exceeding OCSFA alone by 0.9 percentage points. The combined model also achieves the best results on WHU-MARS-1000-GD and WHU-MARS-2337, supporting the complementary benefits of modality-specific spectral alignment and layer-localized cross-view supervision.

As shown in Table 5, we further isolate the effect of OCSFA under Ground-to-Ground retrieval and progressively decompose its transformation. Mean alignment improves Rank-1 across all spectral pairs, indicating the importance of correcting modality-dependent translations. Full OCSFA consistently yields further gains, confirming that mean translation and orthogonal rotation are complementary.

Table 5: Rank-1 for Ground-to-Ground retrieval on WHU-MARS-1000.

Query → Gallery	Before alignment	Mean alignment	Full OCSFA
RGB → NIR	26.36	26.82	28.08
NIR → RGB	23.77	24.83	25.30
RGB → TIR	6.16	6.56	7.95
TIR → RGB	4.83	6.62	9.34
NIR → TIR	4.17	4.70	6.75
TIR → NIR	4.64	6.36	7.62

Appendix E presents sensitivity and transformation-matrix analyses, and additional experimental results.

3.3 Visualizations

Figure 3 shows the rank changes of individual queries after introducing OCSFA and LLCVL. OCSFA increases the number of Rank-1 queries from 1,878 to 2,139, while LLCVL further raises it to 2,200, yielding a Rank-1 accuracy of 34.3%. Dark blue and light orange flows indicate movements toward better and worse ranks, respectively.

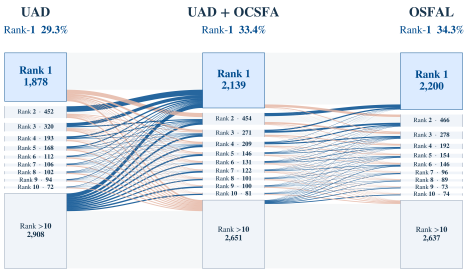

Figure 3: Changes in query retrieval ranks after adding OCSFA and LLCVL to the baseline on WHU-MARS-1000.

To understand LLCVL's additional gain, we examine the block-wise representation dynamics in Figure 2(b-i). Early blocks strengthen identity-discriminative features but may enlarge the Ground–Aerial discrepancy; once identity representations stabilize, deeper blocks begin to reduce the view gap. This pattern underscores the greater difficulty of learning view-invariant representations, consistent with our discussion in Sec. 1. The transition score identifies the boundary where this shift occurs while identity matching remains reliable.

Appendix E presents qualitative retrieval, feature distance density distributions, and additional experimental visualizations.

4 Conclusion

We presented OSFAL for AS-ReID, combining orthogonal spectral alignment with layer-localized cross-view learning. OCSFA centers modality features by means, applies orthogonal rotations, and translates them into a shared reference space while preserving the within-modality geometry. LLCVL uses block-wise statistics to locate a key boundary and applies cross-view contrastive supervision at the selected block output to strengthen identity matching between Ground and Aerial views. Experiments on WHU-MARS demonstrate consistent improvements across multiple protocols.

AI USE STATEMENT

In this work, we used generative AI tools to assist with translation.

We have not used generative AI tools to help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, support qualitative and thematic data analysis, interpret results.

Generating synthetic data sets, cleaning and reformatting dataset are not applicable to this work.

Additionally, we used generative AI tools to edit a research paper to improve readability.

We have reviewed all AI-assisted work. We reviewed and revised all AI-assisted text to ensure technical accuracy and consistency with the conducted research.

We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Our experiments are designed to be fully reproducible. The experimental setup, evaluation protocols, implementation details, and hyperparameter settings are described in the main paper and appendix. Upon acceptance, we will release the complete source code, configuration files, and model checkpoints corresponding to the reported experimental results with the camera-ready version.

REFERENCES

Seokeon Choi, Sumin Lee, Youngeun Kim, Taekyung Kim, and Changick Kim. Hi-cmd: Hierarchical cross-modality disentanglement for visible-infrared person re-identification. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10254–10263. IEEE, 2020.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Jianyang Gu, Kai Wang, Hao Luo, Chen Chen, Wei Jiang, Yuqiang Fang, Shanghang Zhang, Yang You, and Jian Zhao. Msinet: Twins contrastive search of multi-scale interaction for object reid. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19243–19253. IEEE, 2023.

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

Xin Hao, Sanyuan Zhao, Mang Ye, and Jianbing Shen. Cross-modality person re-identification via modality confusion and center aggregation. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16383–16392. IEEE, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 14993–15002. IEEE, 2021.

Weizhen He, Yiheng Deng, Shixiang Tang, Qihao Chen, Qingsong Xie, Yizhou Wang, Lei Bai, Feng Zhu, Rui Zhao, Wanli Ouyang, et al. Instruct-reid: A multi-purpose person re-identification task

with instructions. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17521–17531. IEEE, 2024.

Kyle Helfrich, Devin Willmott, and Qiang Ye. Orthogonal recurrent neural networks with scaled cayley transform. In *International conference on machine learning*, pp. 1969–1978. PMLR, 2018.

Kongzhu Jiang, Tianzhu Zhang, Xiang Liu, Bingqiao Qian, Yongdong Zhang, and Feng Wu. Cross-modality transformer for visible-infrared person re-identification. In *European conference on computer vision*, pp. 480–496. Springer, 2022.

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

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

Minsu Kim, Seungryong Kim, Jungin Park, Seongheon Park, and Kwanghoon Sohn. Partmix: Regularization strategy to learn part discovery for visible-infrared person re-identification. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18621–18632. IEEE, 2023.

Qingming Leng, Mang Ye, and Qi Tian. A survey of open-world person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(4):1092–1108, 2019.

Qiao Li, Jie Li, Yukang Zhang, Lei Tan, Jing Chen, and Jiayi Ji. Gsalign: Geometric and semantic alignment network for aerial-ground person re-identification. *Advances in Neural Information Processing Systems*, 38:93785–93805, 2026.

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

Wen Li, Cheng Zou, Meng Wang, Furong Xu, Jianan Zhao, Ruobing Zheng, Yuan Cheng, and Wei Chu. Dc-former: Diverse and compact transformer for person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 1415–1423, 2023b.

Tengfei Liang, Yi Jin, Zhun Zhong, Xin Chen, Xianjia Meng, Tao Wang, and Yidong Li. M 3-reid: Unifying multi-view, granularity, and modality for video-based visible-infrared person re-identification. *IEEE Transactions on Information Forensics and Security*, 20:13112–13126, 2025.

Guorong Lin, Wei-Shi Zheng, Zuoyong Li, Yao Lu, Xiaowen Ma, and Zhenhua Huang. Vision-language models for person re-identification: a survey and outlook. *Information Fusion*, pp. 104095, 2025.

Yutian Lin, Liang Zheng, Zhedong Zheng, Yu Wu, Zhilan Hu, Chenggang Yan, and Yi Yang. Improving person re-identification by attribute and identity learning. *Pattern recognition*, 95:151–161, 2019.

Haijun Liu, Xiaoheng Tan, and Xichuan Zhou. Parameter sharing exploration and hetero-center triplet loss for visible-thermal person re-identification. *IEEE Transactions on Multimedia*, 23: 4414–4425, 2020.

Hankun Liu, Yujian Zhao, and Guanglin Niu. Try harder: Hard sample generation and learning for cloth-changing person re-id. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 1704–1713, 2025.

Weiyang Liu, Zeju Qiu, Yao Feng, Yuliang Xiu, Yuxuan Xue, Longhui Yu, Haiwen Feng, Zhen Liu, Juyeon Heo, Songyou Peng, et al. Parameter-efficient orthogonal finetuning via butterfly factorization. In *International Conference on Learning Representations*, volume 2024, pp. 38317–38350, 2024.

Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1487–1495. IEEE, 2019.

Hao Luo, Pichao Wang, Yi Xu, Feng Ding, Yanxin Zhou, Fan Wang, Hao Li, and Rong Jin. Self-supervised pre-training for transformer-based person re-identification. *arXiv preprint arXiv:2111.12084*, 2021.

Zhangqiang Ming, Min Zhu, Xiangkun Wang, Jiamin Zhu, Junlong Cheng, Chengrui Gao, Yong Yang, and Xiaoyong Wei. Deep learning-based person re-identification methods: A survey and outlook of recent works. *Image and Vision Computing*, 119:104394, 2022.

Dat Tien Nguyen, Hyung Gil Hong, Ki Wan Kim, and Kang Ryoung Park. Person recognition system based on a combination of body images from visible light and thermal cameras. *Sensors*, 17(3):605, 2017.

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

Enhao Ning, Changshuo Wang, Huang Zhang, Xin Ning, and Prayag Tiwari. Occluded person re-identification with deep learning: a survey and perspectives. *Expert systems with applications*, 239:122419, 2024.

Hyunjong Park, Sanghoon Lee, Junghyup Lee, and Bumsub Ham. Learning by aligning: Visible-infrared person re-identification using cross-modal correspondences. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12026–12035. IEEE, 2021.

Zeju Qiu, Weiyang Liu, Haiwen Feng, Yuxuan Xue, Yao Feng, Zhen Liu, Dan Zhang, Adrian Weller, and Bernhard Schölkopf. Controlling text-to-image diffusion by orthogonal finetuning. *Advances in neural information processing systems*, 36:79320–79362, 2023.

Zeju Qiu, Weiyang Liu, Adrian Weller, and Bernhard Schölkopf. Orthogonal finetuning made scalable. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 31934–31951, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *Advances in neural information processing systems*, 34:12116–12128, 2021.

Kaijie Ren and Lei Zhang. Implicit discriminative knowledge learning for visible-infrared person re-identification. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 393–402. IEEE, 2024.

Jiangming Shi, Xiangbo Yin, Yeyun Chen, Yachao Zhang, Zhizhong Zhang, Yuan Xie, and Yanyun Qu. Multi-memory matching for unsupervised visible-infrared person re-identification. In *European Conference on Computer Vision*, pp. 456–474. Springer, 2024a.

Jiangming Shi, Xiangbo Yin, Yachao Zhang, Zhizhong Zhang, Yuan Xie, and Yanyun Qu. Learning commonality, divergence and variety for unsupervised visible-infrared person re-identification. *Advances in Neural Information Processing Systems*, 37:99715–99734, 2024b.

Xiujun Shu, Xiao Wang, Xianghao Zang, Shiliang Zhang, Yuanqi Chen, Ge Li, and Qi Tian. Large-scale spatio-temporal person re-identification: Algorithms and benchmark. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(7):4390–4403, 2021.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017.

Asher Trockman and J Zico Kolter. Orthogonalizing convolutional layers with the cayley transform. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=Pbj8H_jEHYv.

Guan'an Wang, Tianzhu Zhang, Jian Cheng, Si Liu, Yang Yang, and Zengguang Hou. Rgb-infrared cross-modality person re-identification via joint pixel and feature alignment. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3622–3631. IEEE, 2019.

Jiangming Wang, Zhizhong Zhang, Mingang Chen, Yi Zhang, Cong Wang, Bin Sheng, Yanyun Qu, and Yuan Xie. Optimal transport for label-efficient visible-infrared person re-identification. In *European Conference on Computer Vision*, pp. 93–109. Springer, 2022.

Menglin Wang, Baisheng Lai, Jianqiang Huang, Xiaojin Gong, and Xian-Sheng Hua. Camera-aware proxies for unsupervised person re-identification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 2764–2772, 2021.

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

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning*, pp. 9929–9939. PMLR, 2020.

Yuhai Wang and Maryam Pishgar. Dynamic token selection for aerial-ground person re-identification. *arXiv preprint arXiv:2412.00433*, 2024.

Yuhao Wang, Xuehu Liu, Pingping Zhang, Hu Lu, Zhengzheng Tu, and Huchuan Lu. Top-reid: Multi-spectral object re-identification with token permutation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 5758–5766, 2024.

Yuhao Wang, Xiang Hu, Lixin Wang, Pingping Zhang, and Huchuan Lu. Sd-reid: view-aware stable diffusion for aerial-ground person re-identification. *IEEE Transactions on Image Processing*, 2026.

Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong, and Jianhuang Lai. Rgb-infrared cross-modality person re-identification. In *2017 IEEE international conference on computer vision (ICCV)*, pp. 5390–5399. IEEE, 2017.

Bin Yang, Jun Chen, and Mang Ye. Shallow-deep collaborative learning for unsupervised visible-infrared person re-identification. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16870–16879. IEEE, 2024.

Chaolei Yang, Linhui Dai, Jianlin Xiang, Wenhao Li, and Yanshan Li. Vir-reid: Learning view-invariant representations for aerial-ground person re-identification. *IEEE Transactions on Multimedia*, 2026.

Haoyu Yao, Bin Yang, Wenke Huang, Bo Du, and Mang Ye. Unsupervised visible-infrared person re-identification under unpaired settings. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11916–11926. IEEE, 2025.

Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE transactions on pattern analysis and machine intelligence*, 44(6):2872–2893, 2021.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. *arXiv preprint arXiv:2410.06940*, 2024.

Guiwei Zhang, Yongfei Zhang, Tianyu Zhang, Bo Li, and Shiliang Pu. Pha: Patch-wise high-frequency augmentation for transformer-based person re-identification. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14133–14142. IEEE, 2023.

Qiang Zhang, Changzhou Lai, Jianan Liu, Nianchang Huang, and Jungong Han. Fmcnet: Feature-level modality compensation for visible-infrared person re-identification. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7339–7348. IEEE, 2022.

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

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

Yukang Zhang and Hanzi Wang. Diverse embedding expansion network and low-light cross-modality benchmark for visible-infrared person re-identification. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2153–2162. IEEE, 2023.

Yujian Zhao, Chengru Wu, Yinong Xu, Xuanzheng Du, Ruiyu Li, and Guanglin Niu. Ccup: A controllable synthetic data generation pipeline for pretraining cloth-changing person re-identification models. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6. IEEE, 2025.

Yujian Zhao, Yukang Zhao, Hankun Liu, Haoxuan Xu, Bo Li, Hanzi Wan, and Guanglin Niu. Dual-space modality consistency learning for universal cross-modal re-identification. *arXiv preprint arXiv:2608.06943*, 2026a.

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

Aihua Zheng, Hao Xie, Xixi Wan, Zi Wang, Shihao Li, Jin Tang, and Bin Luo. Semantic-driven visual progressive refinement for aerial-ground person reid: A challenging large-scale benchmark. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 13360–13368, 2026.

Xiangtao Zheng, Xiumei Chen, and Xiaoqiang Lu. Visible-infrared person re-identification via partially interactive collaboration. *IEEE Transactions on Image Processing*, 31:6951–6963, 2022.

Jialong Zuo, Yongtai Deng, Mengdan Tan, Rui Jin, Dongyue Wu, Nong Sang, Liang Pan, and Changxin Gao. Reid5o: Achieving omni multi-modal person re-identification in a single model. *Advances in Neural Information Processing Systems*, 38:53401–53424, 2026.

A Related work

A.1 Any-scenario person re-identification

Person re-identification (Lin et al., 2019; Ye et al., 2021; Leng et al., 2019; Ming et al., 2022; Ning et al., 2024; Shu et al., 2021; Zhao et al., 2025; Liu et al., 2025) has been extended beyond conventional RGB camera networks to accommodate heterogeneous sensing conditions. Visible–infrared benchmarks, such as SYSU-MM01 (Wu et al., 2017) and LLCM (Zhang & Wang, 2023), emphasize spectral discrepancies and illumination changes, whereas aerial–ground benchmarks, including CARGO (Zhang et al., 2024) and LAGPeR (Wang et al., 2025), emphasize viewpoint differences between fixed cameras and aerial platforms. WHU-MARS brings these challenges together by providing RGB, NIR, and TIR observations from both Ground and Aerial views. The associated AS-ReID task requires a single model to retrieve identities from a heterogeneous gallery spanning all available modality–view conditions (Zhao et al., 2026b).

A.2 Cross-spectral feature alignment

Cross-spectral ReID (Zhang et al., 2022; Choi et al., 2020; Ren & Zhang, 2024; Park et al., 2021; Liang et al., 2025; Zhao et al., 2026a) seeks representations that bridge sensor-dependent appearance differences while retaining identity information. Existing approaches explore pixel-level alignment, local correspondence learning, feature augmentation, and knowledge transfer. AlignGAN jointly aligns images and features through adversarial learning to reduce cross-modality discrepancies (Wang et al., 2019). Learning by Aligning (LbA) exploits dense cross-modal correspondences to associate local features and alleviate spatial misalignment (Park et al., 2021). Feature augmentation offers another route: PartMix mixes part descriptors within and across modalities for contrastive learning (Kim et al., 2023), while DEEN generates diverse embeddings to enrich cross-modality feature learning (Zhang & Wang, 2023). At the representation level, MCLNet combines modality confusion with identity-aware center aggregation to encourage modality-invariant and identity-discriminative features (Hao et al., 2021). IDKL further emphasizes the identity information contained in modality-specific features, purifying these features and distilling their discriminative knowledge into modality-shared representations (Ren & Zhang, 2024). Under unsupervised settings, Multi-Memory Matching (MMM) represents pseudo-identities with multiple memories and introduces soft cluster-level alignment to account for intra-class variation and noisy cross-modality correspondences (Shi et al., 2024a). More recently, Mapping and Collaborative Learning (MCL) addresses cross-modality identity mismatches in unpaired training data through feature mapping and collaborative alignment at the cluster and instance levels (Yao et al., 2025). Together, these studies highlight the need to reduce spectral discrepancies while preserving discriminative identity information and establishing reliable cross-modality correspondences.

OCSFA approaches this trade-off through a geometric constraint on the alignment map. Orthogonal transformations have also been explored in model adaptation: OFT preserves pairwise angular relationships between neuron weights, and BOFT introduces a butterfly parameterization for efficient orthogonal finetuning (Qiu et al., 2023; Liu et al., 2024). Unlike these weight-space adaptation methods, OCSFA learns modality-specific transformations of image embeddings into a reference coordinate system. Orthogonal rotations are combined with mean translation, preserving within-modality Euclidean distances before normalization. The transformations are learned jointly with the encoder through identity-wise consistency among modality–view prototypes, without requiring explicit image-level spectral pairing in the loss.

A.3 Cross-view feature alignment

Aerial–ground ReID (Nguyen et al., 2024; Wang & Pishgar, 2024; Wang et al., 2025; 2026; Zheng et al., 2026; Li et al., 2026) requires matching identities despite changes in body orientation, visible regions, scale, and occlusion. Existing approaches address these challenges through attribute-guided learning, view-aware architectures, and geometric alignment. The three-stream network accompanying AG-ReID.v2 combines attribute-guided attention with head-region features to strengthen identity representations under elevated viewpoints (Nguyen et al., 2024). VDT separates view-related and view-unrelated features using hierarchical subtractive separation and an orthogonal loss (Zhang et al., 2024). Its orthogonality constraint concerns the relationship between feature components, rather than an isometric transformation of embeddings. DTST selects informative tokens to strengthen identity representations (Wang & Pishgar, 2024), while SeCap combines input-adaptive prompt recalibration with local feature refinement (Wang et al., 2025). Beyond architectural design, VIF-AGReID combines patch-level rotational augmentation with a view-invariant angular loss to reduce sensitivity to viewpoint changes (Khalid et al., 2025). To address spatial misalignment more explicitly, GSAlign uses learnable thin-plate-spline transformations to warp pedestrian features and visibility-aware masks to emphasize observable body regions (Li et al., 2026). More recently, ViSA uses view-aware experts to generate semantic queries and graph reasoning to align local regions across views (Zhang et al., 2026). SD-ReID explores a generative alternative, using identity- and view-conditioned Stable Diffusion to model features across viewpoints and combining these features with person representations for retrieval (Wang et al., 2026).

Complementary to architectural design, the depth at which alignment supervision is introduced also merits consideration. In visible–infrared ReID, SDCL exploits complementary shallow and deep features through collaborative contrastive learning (Yang et al., 2024). LLCVL instead examines block-wise identity matching reliability and within-view versus cross-view similarities to locate a

boundary for aerial–ground supervision. It applies bidirectional cross-view contrastive learning at the selected block output within each spectral modality. The distinction lies in selecting the supervision location from representation statistics, without introducing an additional inference branch.

B Detailed algebraic properties of orthogonal transformations

This appendix provides detailed derivations of the isometric properties of orthogonal transformations and the Cayley parameterization used to construct the full-dimensional orthogonal maps in our method.

B.1 Orthogonal Transformations Are Isometries

Let $\mathbf{R} \in \mathbb{R}^{d\times d}$ be an orthogonal matrix satisfying

$$\mathbf{R}^\top \mathbf{R} = \mathbf{I}_d. \tag{20}$$

Since $\mathbf{R}$ is square, Equation 20 implies that $\mathbf{R}^\top$ is its inverse:

$$\mathbf{R}^{-1} = \mathbf{R}^\top. \tag{21}$$

We next show that applying $\mathbf{R}$ preserves the geometric relations between feature vectors.

Norm preservation. For any $\mathbf{z}_i \in \mathbb{R}^d$, the squared norm of its transformed representation is

$$\begin{aligned} \|\mathbf{R}\mathbf{z}_i\|_2^2 &= (\mathbf{R}\mathbf{z}_i)^\top (\mathbf{R}\mathbf{z}_i) \\ &= \mathbf{z}_i^\top \mathbf{R}^\top \mathbf{R}\mathbf{z}_i \\ &= \mathbf{z}_i^\top \mathbf{I}_d \mathbf{z}_i \\ &= \mathbf{z}_i^\top \mathbf{z}_i \\ &= \|\mathbf{z}_i\|_2^2 . \end{aligned} \tag{22}$$

Both sides are nonnegative, so taking their square roots gives

$$\|\mathbf{R}\mathbf{z}_i\|_2 = \|\mathbf{z}_i\|_2 . \tag{23}$$

Inner-product preservation. For any two vectors $\mathbf{z}_i, \mathbf{z}_j \in \mathbb{R}^d$, we have

$$\begin{aligned} (\mathbf{R}\mathbf{z}_i)^\top (\mathbf{R}\mathbf{z}_j) &= \mathbf{z}_i^\top \mathbf{R}^\top \mathbf{R}\mathbf{z}_j \\ &= \mathbf{z}_i^\top \mathbf{I}_d \mathbf{z}_j \\ &= \mathbf{z}_i^\top \mathbf{z}_j . \end{aligned} \tag{24}$$

Euclidean-distance preservation. The difference between two transformed vectors can be written as

$$\mathbf{R}\mathbf{z}_i - \mathbf{R}\mathbf{z}_j = \mathbf{R}(\mathbf{z}_i - \mathbf{z}_j). \tag{25}$$

Using Equation 23, its Euclidean distance is therefore

$$\begin{aligned} \|\mathbf{R}\mathbf{z}_i - \mathbf{R}\mathbf{z}_j\|_2 &= \|\mathbf{R}(\mathbf{z}_i - \mathbf{z}_j)\|_2 \\ &= \|\mathbf{z}_i - \mathbf{z}_j\|_2 . \end{aligned} \tag{26}$$

Cosine-similarity preservation. For nonzero vectors $\mathbf{z}_i$ and $\mathbf{z}_j$, their cosine similarity after transformation is

$$\begin{aligned} \mathrm{sim}_{\cos}(\mathbf{R}\mathbf{z}_i, \mathbf{R}\mathbf{z}_j) &= \frac{(\mathbf{R}\mathbf{z}_i)^\top (\mathbf{R}\mathbf{z}_j)}{\|\mathbf{R}\mathbf{z}_i\|_2 \|\mathbf{R}\mathbf{z}_j\|_2} \\ &= \frac{\mathbf{z}_i^\top \mathbf{z}_j}{\|\mathbf{z}_i\|_2 \|\mathbf{z}_j\|_2} \\ &= \mathrm{sim}_{\cos}(\mathbf{z}_i, \mathbf{z}_j), \end{aligned} \tag{27}$$

where the second equality follows from Equations 23 and 24.

Equations 26 and 27 show that an orthogonal transformation preserves both Euclidean-distance and cosine-similarity rankings when the same transformation is applied to all features in a domain. It therefore changes the coordinate orientation without scaling, shearing, or collapsing the original feature geometry.

B.2 FULL-DIMENSIONAL ORTHOGONAL MAPS VIA THE CAYLEY TRANSFORM

Construction of a skew-symmetric generator. The implementation begins with an unconstrained trainable matrix $\mathbf{W} \in \mathbb{R}^{d\times d}$. We construct

$$\mathbf{A} = \frac{1}{2}\left(\mathbf{W} - \mathbf{W}^\top\right). \tag{28}$$

Taking its transpose gives

$$\begin{aligned}\mathbf{A}^\top &= \frac{1}{2}\left(\mathbf{W} - \mathbf{W}^\top\right)^\top \\ &= \frac{1}{2}\left(\mathbf{W}^\top - \mathbf{W}\right) \\ &= -\frac{1}{2}\left(\mathbf{W} - \mathbf{W}^\top\right) \\ &= -\mathbf{A}.\end{aligned} \tag{29}$$

Thus, $\mathbf{A}$ is skew-symmetric for any value of the unconstrained parameter $\mathbf{W}$.

Invertibility of $\mathbf{I}_d + \mathbf{A}$. The Cayley transform requires $\mathbf{I}_d + \mathbf{A}$ to be invertible. Suppose that a vector $\mathbf{u} \in \mathbb{R}^d$ satisfies

$$(\mathbf{I}_d + \mathbf{A})\mathbf{u} = \mathbf{0}. \tag{30}$$

Multiplying from the left by $\mathbf{u}^\top$ yields

$$\begin{aligned}0 &= \mathbf{u}^\top(\mathbf{I}_d + \mathbf{A})\mathbf{u} \\ &= \mathbf{u}^\top\mathbf{u} + \mathbf{u}^\top\mathbf{A}\mathbf{u} \\ &= \|\mathbf{u}\|_2^2 + \mathbf{u}^\top\mathbf{A}\mathbf{u}.\end{aligned} \tag{31}$$

The scalar $\mathbf{u}^\top\mathbf{A}\mathbf{u}$ satisfies

$$\begin{aligned}\mathbf{u}^\top\mathbf{A}\mathbf{u} &= \left(\mathbf{u}^\top\mathbf{A}\mathbf{u}\right)^\top \\ &= \mathbf{u}^\top\mathbf{A}^\top\mathbf{u} \\ &= -\mathbf{u}^\top\mathbf{A}\mathbf{u}.\end{aligned} \tag{32}$$

It must therefore be zero. Substituting this result into Equation 31 gives

$$\|\mathbf{u}\|_2^2 = 0, \tag{33}$$

which implies $\mathbf{u} = \mathbf{0}$. Hence, $\mathbf{I}_d + \mathbf{A}$ has a trivial null space and is invertible. Applying the same argument to $-\mathbf{A}$ shows that $\mathbf{I}_d - \mathbf{A}$ is also invertible.

Cayley construction. We define the orthogonal transformation as

$$\mathbf{R} = \mathrm{Cay}(\mathbf{A}) = (\mathbf{I}_d + \mathbf{A})^{-1}(\mathbf{I}_d - \mathbf{A}). \tag{34}$$

For compactness, let

$$\mathbf{P} = \mathbf{I}_d + \mathbf{A}, \qquad \mathbf{N} = \mathbf{I}_d - \mathbf{A}. \tag{35}$$

Both matrices are polynomials in $\mathbf{A}$, and therefore commute:

$$\begin{aligned}\mathbf{P}\mathbf{N} &= (\mathbf{I}_d + \mathbf{A})(\mathbf{I}_d - \mathbf{A}) \\ &= \mathbf{I}_d - \mathbf{A}^2 \\ &= (\mathbf{I}_d - \mathbf{A})(\mathbf{I}_d + \mathbf{A}) \\ &= \mathbf{N}\mathbf{P}.\end{aligned} \tag{36}$$

Consequently, their inverses also commute with $\mathbf{P}$ and $\mathbf{N}$.

Using $\mathbf{A}^\top = -\mathbf{A}$, we obtain

$$\mathbf{P}^\top = \mathbf{N}, \qquad \mathbf{N}^\top = \mathbf{P}. \tag{37}$$

Taking the transpose of Equation 34 gives

$$\begin{aligned}\mathbf{R}^\top &= \left(\mathbf{P}^{-1}\mathbf{N}\right)^\top \\ &= \mathbf{N}^\top \left(\mathbf{P}^{-1}\right)^\top \\ &= \mathbf{P}\left(\mathbf{P}^\top\right)^{-1} \\ &= \mathbf{P}\mathbf{N}^{-1}. \end{aligned} \tag{38}$$

Therefore,

$$\begin{aligned}\mathbf{R}^\top\mathbf{R} &= \left(\mathbf{P}\mathbf{N}^{-1}\right)\left(\mathbf{P}^{-1}\mathbf{N}\right) \\ &= \mathbf{P}\mathbf{P}^{-1}\mathbf{N}^{-1}\mathbf{N} \\ &= \mathbf{I}_d. \end{aligned} \tag{39}$$

The reordering in the second line is valid because the matrices commute, as established in Equation 36. Hence, the Cayley transform produces an exactly orthogonal matrix in exact arithmetic.

Full-rank property. Suppose that $\mathbf{R}\mathbf{u} = \mathbf{0}$. Multiplying by $\mathbf{R}^\top$ gives

$$\begin{aligned}\mathbf{u} &= \mathbf{I}_d\mathbf{u} \\ &= \mathbf{R}^\top\mathbf{R}\mathbf{u} \\ &= \mathbf{R}^\top\mathbf{0} \\ &= \mathbf{0}. \end{aligned} \tag{40}$$

Thus, $\mathbf{R}$ has a trivial null space. Since $\mathbf{R} \in \mathbb{R}^{d\times d}$ is square,

$$\operatorname{rank}(\mathbf{R}) = d. \tag{41}$$

Notice that this conclusion concerns the rank of $\mathbf{R}$; the skew-symmetric generator $\mathbf{A}$ itself need not be full rank.

A $d \times d$ skew-symmetric matrix has zero diagonal entries. Each entry above the diagonal uniquely determines the corresponding entry below the diagonal:

$$A_{ji} = -A_{ij}. \tag{42}$$

There are

$$\frac{d(d-1)}{2} \tag{43}$$

entries above the diagonal. The dense Cayley parameterization therefore provides $d(d-1)/2$ effective degrees of freedom and does not impose a predefined low-rank factorization on the resulting orthogonal map.

Identity initialization. When the unconstrained parameter is initialized as $\mathbf{W} = \mathbf{0}$, Equation 28 gives $\mathbf{A} = \mathbf{0}$. Equation 34 then gives

$$\begin{aligned}\mathbf{R} &= (\mathbf{I}_d + \mathbf{0})^{-1}(\mathbf{I}_d - \mathbf{0}) \\ &= \mathbf{I}_d^{-1}\mathbf{I}_d \\ &= \mathbf{I}_d. \end{aligned} \tag{44}$$

The transformation therefore begins as an identity map and is optimized continuously with the rest of the network.

B.3 Scope of the Cayley parameterization

The Cayley transform does not represent orthogonal matrices with eigenvalue -1. To see this, adding $\mathbf{I}_d$ to Equation 34 gives

$$\begin{aligned}\mathbf{I}_d + \mathbf{R} &= \mathbf{I}_d + (\mathbf{I}_d + \mathbf{A})^{-1}(\mathbf{I}_d - \mathbf{A}) \\ &= (\mathbf{I}_d + \mathbf{A})^{-1}\left[(\mathbf{I}_d + \mathbf{A}) + (\mathbf{I}_d - \mathbf{A})\right] \\ &= 2(\mathbf{I}_d + \mathbf{A})^{-1}. \end{aligned} \tag{45}$$

The right-hand side is invertible, so $\mathbf{I}_d + \mathbf{R}$ is invertible. Therefore, no nonzero vector $\mathbf{u}$ can satisfy $\mathbf{R}\mathbf{u} = -\mathbf{u}$, and -1 cannot be an eigenvalue of $\mathbf{R}$.

Conversely, let $\mathbf{R}$ be an orthogonal matrix for which $\mathbf{I}_d + \mathbf{R}$ is invertible. Define

$$\mathbf{A} = (\mathbf{I}_d - \mathbf{R})(\mathbf{I}_d + \mathbf{R})^{-1}. \tag{46}$$

Taking the transpose yields

$$\mathbf{A}^\top = (\mathbf{I}_d + \mathbf{R}^\top)^{-1}(\mathbf{I}_d - \mathbf{R}^\top). \tag{47}$$

Since $\mathbf{R}^\top = \mathbf{R}^{-1}$, we have

$$\mathbf{I}_d + \mathbf{R}^\top = \mathbf{R}^\top(\mathbf{I}_d + \mathbf{R}), \tag{48}$$

$$\mathbf{I}_d - \mathbf{R}^\top = \mathbf{R}^\top(\mathbf{R} - \mathbf{I}_d). \tag{49}$$

Substituting these identities into Equation 47 gives

$$\begin{aligned}
\mathbf{A}^\top &= \left[\mathbf{R}^\top(\mathbf{I}_d + \mathbf{R})\right]^{-1}\mathbf{R}^\top(\mathbf{R} - \mathbf{I}_d) \\
&= (\mathbf{I}_d + \mathbf{R})^{-1}\mathbf{R}\mathbf{R}^\top(\mathbf{R} - \mathbf{I}_d) \\
&= (\mathbf{I}_d + \mathbf{R})^{-1}(\mathbf{R} - \mathbf{I}_d) \\
&= -(\mathbf{I}_d - \mathbf{R})(\mathbf{I}_d + \mathbf{R})^{-1} \\
&= -\mathbf{A}.
\end{aligned} \tag{50}$$

Thus, the inverse Cayley transform produces a skew-symmetric matrix.

Finally,

$$\mathbf{I}_d + \mathbf{A} = 2(\mathbf{I}_d + \mathbf{R})^{-1}, \tag{51}$$

$$\mathbf{I}_d - \mathbf{A} = 2\mathbf{R}(\mathbf{I}_d + \mathbf{R})^{-1}. \tag{52}$$

Substituting them into the Cayley transform gives

$$\begin{aligned}
\mathrm{Cay}(\mathbf{A}) &= (\mathbf{I}_d + \mathbf{A})^{-1}(\mathbf{I}_d - \mathbf{A}) \\
&= \left[2(\mathbf{I}_d + \mathbf{R})^{-1}\right]^{-1}\left[2\mathbf{R}(\mathbf{I}_d + \mathbf{R})^{-1}\right] \\
&= (\mathbf{I}_d + \mathbf{R})\mathbf{R}(\mathbf{I}_d + \mathbf{R})^{-1} \\
&= \mathbf{R},
\end{aligned} \tag{53}$$

where the last equality follows because $\mathbf{R}$ commutes with $\mathbf{I}_d + \mathbf{R}$. The Cayley transform therefore parameterizes precisely the orthogonal matrices without eigenvalue -1, while always producing a full-rank orthogonal map.

C Additional Details of LLCVL Localization

Epoch-wise accumulation and split construction. At the beginning of each epoch, we initialize an empty feature-statistics bank for every block and identity–modality–view cell. Suppose the samples of a cell $\mathcal{C}_{y,m,v}$ are observed in the order $i_1, i_2, \ldots, i_N$. We assign them alternately to two disjoint subsets:

$$\begin{aligned}
\mathcal{C}^{(1)}_{y,m,v} &= \{i_j \in \mathcal{C}_{y,m,v} \mid j \text{ is odd}\}, \\
\mathcal{C}^{(2)}_{y,m,v} &= \{i_j \in \mathcal{C}_{y,m,v} \mid j \text{ is even}\}.
\end{aligned} \tag{54}$$

This deterministic interleaving prevents either subset from being formed exclusively from an early or late portion of the epoch.

For each block ℓ and split $r \in \{1, 2\}$, we accumulate the sufficient statistics

$$\mathbf{a}^{(\ell)}_{y,m,v,r} = \sum_{i \in \mathcal{C}^{(r)}_{y,m,v}} \overline{\mathbf{z}}^{(\ell)}_i, \qquad n_{y,m,v,r} = \left|\mathcal{C}^{(r)}_{y,m,v}\right|. \tag{55}$$

In distributed training, these sums and counts are aggregated across all workers at the end of the epoch. The full-view prototype used for cross-view comparison is computed from all observations in the cell:

$$\mathbf{p}^{(\ell)}_{y,m,v} = \mathrm{norm}\left(\frac{\mathbf{a}^{(\ell)}_{y,m,v,1} + \mathbf{a}^{(\ell)}_{y,m,v,2}}{n_{y,m,v,1} + n_{y,m,v,2}}\right). \tag{56}$$

By contrast, the two split prototypes used to estimate within-view stability are

$$\mathbf{p}^{(\ell,r)}_{y,m,v} = \mathrm{norm}\left(\frac{\mathbf{a}^{(\ell)}_{y,m,v,r}}{n_{y,m,v,r}}\right), \qquad r \in \{1,2\}. \tag{57}$$

Thus, Equation 56 is a sample-count-weighted estimate based on the complete cell and is not obtained by averaging the two normalized split prototypes. The split is introduced solely to provide two independent estimates of the same view condition.

Identity-partition voting. To reduce the sensitivity of the selected boundary to a particular set of identities, we evaluate the transition statistics over multiple fixed identity partitions. For each random seed $\xi \in \{1234, 42, 2026\}$, the training identities are randomly permuted and divided into $K = 5$ disjoint folds. Let $\mathcal{F}_{\xi,k}$ denote fold k under seed ξ. For modality m, only identities having nonempty split subsets under both views are retained:

$$\mathcal{Y}_{\xi,k,m} = \{y \in \mathcal{F}_{\xi,k} \mid n_{y,m,v,r} > 0,\ \forall v \in \{\mathrm{G},\mathrm{A}\},\ \forall r \in \{1,2\}\}. \tag{58}$$

A fold–modality combination is considered valid when $|\mathcal{Y}_{\xi,k,m}| \geq 2$. We compute $\kappa^{(\ell)}_{m,\xi,k}$, $\delta^{(\ell)}_{m,\xi,k}$, and $\tau^{(\ell)}_{m,\xi,k}$ using only these identities. It then casts a vote for the boundary

$$b_{\xi,k,m} = \operatorname*{arg\,max}_{\ell \in \{1,\ldots,L-1\}} \tau^{(\ell)}_{m,\xi,k}. \tag{59}$$

We additionally obtain a modality-aggregated decision from every fold for which all modalities are valid. Specifically, we first average the reliability and view discrepancy across modalities,

$$\overline{\kappa}^{(\ell)}_{\xi,k} = \frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}} \kappa^{(\ell)}_{m,\xi,k}, \qquad \overline{\delta}^{(\ell)}_{\xi,k} = \frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}} \delta^{(\ell)}_{m,\xi,k}, \tag{60}$$

and apply Equation 14 to these averaged statistics to obtain $\overline{\tau}^{(\ell)}_{\xi,k}$. The corresponding macro-level vote is

$$b^{\mathrm{macro}}_{\xi,k} = \operatorname*{arg\,max}_{\ell \in \{1,\ldots,L-1\}} \overline{\tau}^{(\ell)}_{\xi,k}. \tag{61}$$

Let $V_{\mathrm{macro}}(\ell)$ and $V_m(\ell)$ denote the numbers of macro-level and modality-specific votes received by boundary ℓ, respectively. The candidate boundary is determined by

$$\ell_{\mathrm{cand}} = \operatorname*{arg\,max}_{\ell \in \{1,\ldots,L-1\}} V_{\mathrm{macro}}(\ell). \tag{62}$$

Ties are resolved deterministically in favor of the shallower boundary.

Consensus gate and activation. Let N_{macro} be the number of valid macro-level decisions, N_m the number of valid decisions for modality m, and $N_{\mathrm{mod}} = \sum_{m\in\mathcal{M}} N_m$. The candidate passes the consensus gate only if

$$\begin{aligned} V_{\mathrm{macro}}(\ell_{\mathrm{cand}}) &\geq \lceil 0.8 N_{\mathrm{macro}} \rceil, \\ V_m(\ell_{\mathrm{cand}}) &\geq \left\lceil \tfrac{2}{3} N_m \right\rceil, \qquad \forall m \in \mathcal{M}, \\ \sum_{m\in\mathcal{M}} V_m(\ell_{\mathrm{cand}}) &\geq \lceil 0.8 N_{\mathrm{mod}} \rceil. \end{aligned} \tag{63}$$

With three partition seeds and five folds, a complete epoch produces $N_{\mathrm{macro}} = 15$ macro decisions and $N_m = 15$ decisions per modality. The corresponding thresholds are therefore 12 macro votes, 10 votes for every modality, and 36 modality-specific votes in total.

Boundary scoring begins at epoch 5. A boundary is accepted only when the same candidate passes the consensus gate for three consecutive epochs. If this condition is satisfied after epoch e, the selected boundary $\ell^\star$ is activated from epoch $e+1$. Cross-view supervision is then applied to $\mathbf{z}^{(\ell^\star)}$, namely the output of block $\ell^\star$ and the input to block $\ell^\star+1$. Before activation, the LLCVL loss weight is zero. After activation, it is linearly increased over five epochs according to

$$\gamma_e = \min\left(1,\ \frac{e - e_{\mathrm{act}} + 1}{5}\right), \qquad e \geq e_{\mathrm{act}}, \tag{64}$$

where e_{act} is the activation epoch and $\lambda_{\mathrm{V}}\gamma_e$ is the effective LLCVL weight. If no candidate satisfies the stability criterion by epoch 30, boundary selection is terminated and LLCVL remains disabled. All localization statistics are detached from the computational graph, so this selection procedure does not directly optimize the encoder.

D Expansion of the Bidirectional LLCVL Objective

For modality m at iteration t, let $N_m^{(t)} = |\mathcal{Y}_m^{(t)}|$ and arrange the valid identities as $\mathcal{Y}_m^{(t)} = \{y_1, \ldots, y_{N_m^{(t)}}\}$. The Ground and Aerial prototype collections follow the same identity ordering. For brevity, let

$$C_{jk} = \mathbf{C}_m^{(t)}(y_j, y_k) = \left(\mathbf{g}_{y_j,m}^{(t)}\right)^\top \mathbf{a}_{y_k,m}^{(t)}. \tag{65}$$

Because both prototypes are ℓ_2-normalized, C_{jk} is their cosine similarity. The diagonal entry C_{jj} compares Ground and Aerial prototypes of the same identity y_j, whereas C_{jk} with $j \neq k$ compares different identities.

For Ground-to-Aerial retrieval, each Ground prototype $\mathbf{g}_{y_j,m}^{(t)}$ is treated as a query and all Aerial prototypes are treated as candidates. The probability assigned to Aerial identity y_k is

$$P_m^{\mathrm{G}\to\mathrm{A}}(y_k \mid y_j) = \frac{\exp(C_{jk}/T)}{\sum_{r=1}^{N_m^{(t)}} \exp(C_{jr}/T)}. \tag{66}$$

The correct candidate has $k = j$, yielding

$$\mathcal{L}_m^{(t,\mathrm{G}\to\mathrm{A})} = -\frac{1}{N_m^{(t)}} \sum_{j=1}^{N_m^{(t)}} \log \frac{\exp(C_{jj}/T)}{\sum_{k=1}^{N_m^{(t)}} \exp(C_{jk}/T)}. \tag{67}$$

Thus, the first cross-entropy term in Equation 17 performs classification over each row of $\mathbf{C}_m^{(t)}$.

For the reverse direction, each Aerial prototype is treated as a query and the Ground prototypes become its candidates. The corresponding probability is

$$P_m^{\mathrm{A}\to\mathrm{G}}(y_k \mid y_j) = \frac{\exp(C_{kj}/T)}{\sum_{r=1}^{N_m^{(t)}} \exp(C_{rj}/T)}, \tag{68}$$

and the Aerial-to-Ground loss is

$$\mathcal{L}_m^{(t,\mathrm{A}\to\mathrm{G})} = -\frac{1}{N_m^{(t)}} \sum_{j=1}^{N_m^{(t)}} \log \frac{\exp(C_{jj}/T)}{\sum_{k=1}^{N_m^{(t)}} \exp(C_{kj}/T)}. \tag{69}$$

This is equivalent to applying row-wise cross-entropy to $(\mathbf{C}_m^{(t)})^\top/T$. Consequently,

$$\mathcal{L}_m^{(t)} = \frac{1}{2}\left(\mathcal{L}_m^{(t,\mathrm{G}\to\mathrm{A})} + \mathcal{L}_m^{(t,\mathrm{A}\to\mathrm{G})}\right). \tag{70}$$

Minimizing this objective increases the diagonal same-identity similarities relative to the off-diagonal cross-identity similarities in both retrieval directions. A modality contributes to the objective only when $N_m^{(t)} \geq 2$, ensuring that every query has one positive identity and at least one cross-identity negative.

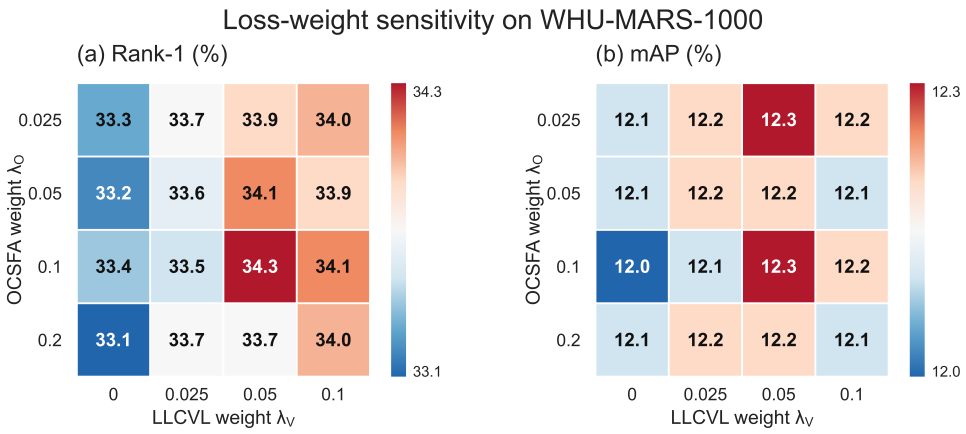

Figure 4: Sensitivity of Rank-1 and mAP to the OCSFA and LLCVL loss weights on WHU-MARS-1000. Rows correspond to the OCSFA coefficient λ_{O}, and columns correspond to the LLCVL coefficient λ_{V}. Blue and red indicate lower and higher performance, respectively.

E Additional Experimental Results

E.1 Hyperparameter sensitivity analysis

We analyze the sensitivity of OSFAL to the coefficients of its two auxiliary objectives. Specifically, λ_{O} controls the OCSFA consistency loss, while λ_{V} controls the LLCVL cross-view contrastive loss. We vary $\lambda_{\mathrm{O}} \in \{0.025, 0.05, 0.10, 0.20\}$ and $\lambda_{\mathrm{V}} \in \{0, 0.025, 0.05, 0.10\}$ on WHU-MARS-1000, while keeping the remaining training and evaluation settings unchanged.

Figure 4 shows that OSFAL is relatively stable across the evaluated coefficient combinations. Rank-1 varies from 33.1% to 34.3%, while mAP remains between 12.0% and 12.3%. When $\lambda_{\mathrm{O}} = 0.10$, increasing λ_{V} from 0 to 0.05 improves Rank-1 from 33.4% to 34.3% and mAP from 12.0% to 12.3%. A further increase to $\lambda_{\mathrm{V}} = 0.10$ slightly reduces the performance, suggesting that an excessively strong cross-view objective may interfere with the overall identity-discriminative representation.

Similarly, with $\lambda_{\mathrm{V}} = 0.05$, increasing λ_{O} from 0.025 to 0.10 raises Rank-1 from 33.9% to 34.3%. Increasing it further to 0.20 reduces Rank-1 to 33.7%, indicating that moderate spectral consistency is preferable to over-regularizing the canonicalized representations. The default setting $(\lambda_{\mathrm{O}}, \lambda_{\mathrm{V}}) = (0.10, 0.05)$ achieves the highest Rank-1 of 34.3% and a joint-highest mAP of 12.3%. We therefore use this coefficient combination in all main experiments.

Overall, OSFAL is not highly sensitive to either coefficient and remains robust to moderate hyper-parameter perturbations, which cause only limited performance degradation.

E.2 Modality-pair retrieval results

We further evaluate all nine query–gallery combinations of RGB, NIR, and TIR on WHU-MARS-1000. For each combination, the query and gallery sets are restricted to the specified spectral modalities, while retaining both Ground and Aerial observations. Thus, same-spectral retrieval still involves view variation. Both methods use their fixed AS-ReID-trained Epoch-120 checkpoints without additional training.

As shown in Table 6, OSFAL improves Rank-1 in seven of the nine combinations, mAP in eight, and mINP in all nine. The largest Rank-1 gains occur in same-spectral retrieval; in particular, NIR-to-NIR Rank-1 increases from 27.82% to 33.40%. Cross-spectral improvements vary by direction: all three metrics improve for RGB-to-NIR, NIR-to-RGB, and RGB-to-TIR retrieval. However, Rank-1 decreases from 6.98% to 6.60% for NIR-to-TIR and from 8.24% to 7.17% for TIR-to-NIR, despite improvements in mAP and mINP. TIR-to-RGB mAP also decreases slightly, from 5.89% to 5.86%.

Table 6: Modality-pair retrieval results (%) on WHU-MARS-1000. UAD denotes our reproduced baseline. The better result for each metric within each query–gallery combination is bold.

Query → Gallery	UAD (Baseline)			OSFAL (Ours)		
	R-1	mAP	mINP	R-1	mAP	mINP
RGB → RGB	40.80	26.37	5.01	**45.39**	**30.30**	**6.22**
RGB → NIR	26.23	15.96	2.28	**27.17**	**17.33**	**2.51**
RGB → TIR	6.89	5.59	1.12	**7.63**	**5.88**	**1.13**
NIR → RGB	24.64	15.62	2.85	**25.57**	**16.96**	**3.16**
NIR → NIR	27.82	16.75	2.44	**33.40**	**19.92**	**2.90**
NIR → TIR	**6.98**	5.26	1.04	6.60	**5.38**	**1.11**
TIR → RGB	7.73	**5.89**	1.31	**8.52**	5.86	**1.35**
TIR → NIR	**8.24**	5.25	1.00	7.17	**5.33**	**1.05**
TIR → TIR	20.66	13.56	2.33	**24.92**	**15.81**	**2.62**

E.3 Analysis of learned orthogonal transformations

We analyze the modality transformations learned by OSFAL on WHU-MARS-1000 using the Epoch-120 checkpoint. The feature dimension is $d = 768$, and RGB is the reference modality. Our analysis addresses three questions: how strongly the learned NIR and TIR maps rotate the feature space, where these adjustments are concentrated, and which properties of the feature distribution are preserved by canonicalization.

E.3.1 Rotation spectrum

A two-dimensional rotation is described by a single angle. In higher dimensions, a rotation can be decomposed into rotations acting in mutually orthogonal two-dimensional planes, with potentially different angles in different planes. The rotation spectrum collects these angles and describes how strongly the transformation acts across the feature space.

The learned map is constructed using the Cayley transform:

$$\mathbf{R}_m = (\mathbf{I}_d + \mathbf{A}_m)^{-1}(\mathbf{I}_d - \mathbf{A}_m), \qquad \mathbf{A}_m^\top = -\mathbf{A}_m. \tag{71}$$

Since $\mathbf{A}_m$ is real and skew-symmetric, its eigenvalues are purely imaginary. Let $\mathrm{i}a_{m,j}$ be one such eigenvalue, where $a_{m,j} \in \mathbb{R}$ and $\mathrm{i}^2 = -1$, and let $\mathbf{v}$ be a corresponding complex eigenvector. Then

$$(\mathbf{I}_d \pm \mathbf{A}_m)\mathbf{v} = (1 \pm \mathrm{i}a_{m,j})\mathbf{v}. \tag{72}$$

The corresponding eigenvalue of $\mathbf{R}_m$ is therefore

$$\lambda_j(\mathbf{R}_m) = \frac{1 - \mathrm{i}a_{m,j}}{1 + \mathrm{i}a_{m,j}} = \exp(-2\mathrm{i}\arctan a_{m,j})\,. \tag{73}$$

Its magnitude is one, so every eigenvalue lies on the unit circle. Nonreal eigenvalues occur in conjugate pairs, with each pair describing a rotation in a two-dimensional plane. The absolute rotation angles are

$$\vartheta_{m,j} = 2\arctan|a_{m,j}| \in [0, \pi). \tag{74}$$

Each rotation plane contributes two equal absolute angles to the full spectrum. The conjugate eigenvalues have unit product, and any real eigenvalues are $+1$. Consequently, $\det(\mathbf{R}_m) = 1$: this parameterization produces rotations without reflections. An eigenvalue of -1, corresponding to an angle of π, is not attained at finite parameter values (Helfrich et al., 2018).

For numerical analysis, we recover the angles from the singular values $\sigma_{m,j}$ of $\mathbf{R}_m - \mathbf{I}_d$ using $\vartheta_{m,j} = 2\arcsin(\sigma_{m,j}/2)$, as derived below. Figure 5(a) plots the absolute angles in descending order. The NIR and TIR maps have mean angles of $13.25°$ and $18.37°$, with maximum angles of $61.86°$ and $83.25°$, respectively.

These results show that the learned TIR map makes larger rotational adjustments on average, while both maps apply stronger adjustments to some subspaces than to others. The reported angles characterize the matrices themselves. They are not the average angular changes of individual sample features, which also depend on how those features are distributed across the rotation planes.

E.3.2 CONCENTRATION OF TRANSFORMATION ENERGY

The rotation spectrum describes the strength of the adjustments. We next examine whether these adjustments are spread across many directions or concentrated in a smaller subset.

Every orthogonal matrix has unit singular values, so the singular values of $\mathbf{R}_m$ itself cannot distinguish a weak rotation from a strong one. Instead, we analyze

$$\mathbf{D}_m = \mathbf{R}_m - \mathbf{I}_d. \tag{75}$$

The identity matrix leaves every feature unchanged. Thus, $\mathbf{D}_m$ measures the change introduced by the learned rotation.

Using $\mathbf{R}_m^\top \mathbf{R}_m = \mathbf{I}_d$, we obtain

$$\begin{aligned} \mathbf{D}_m^\top \mathbf{D}_m &= (\mathbf{R}_m^\top - \mathbf{I}_d)(\mathbf{R}_m - \mathbf{I}_d) \\ &= 2\mathbf{I}_d - \mathbf{R}_m - \mathbf{R}_m^\top. \end{aligned} \tag{76}$$

For a rotation angle $\vartheta_{m,j}$, the corresponding eigenvalue of this matrix is

$$2 - 2\cos\vartheta_{m,j} = 4\sin^2(\vartheta_{m,j}/2). \tag{77}$$

The singular values of $\mathbf{D}_m$ therefore satisfy

$$\sigma_{m,j} = 2\sin(\vartheta_{m,j}/2), \qquad \vartheta_{m,j} = 2\arcsin(\sigma_{m,j}/2). \tag{78}$$

Larger singular values correspond to directions with larger changes.

After sorting the singular values in descending order, we define

$$E_m(r) = \frac{\sum_{j=1}^{r} \sigma_{m,j}^2}{\sum_{j=1}^{d} \sigma_{m,j}^2}, \qquad r_{m,0.90} = \min\{r : E_m(r) \geq 0.90\}. \tag{79}$$

Here, $E_m(r)$ measures the proportion of the total squared matrix change accounted for by the leading r singular directions. Equivalently, $1 - E_m(r)$ is the relative squared Frobenius error of the best rank-r approximation to $\mathbf{D}_m$.

As shown in Figure 5(b), 248 directions for NIR and 240 for TIR account for 90% of their respective transformation energies. Thus, approximately one third of the 768 directions account for most of the transformation, while the remaining directions contribute smaller adjustments. Both displacement matrices nevertheless have numerical rank 768 at a singular-value threshold of 10^{-6}.

These singular directions are combinations of the original feature coordinates, rather than a selected subset of feature dimensions. The energy percentage describes the matrix change; it does not measure the proportion of identity information or retrieval performance retained. The normalized orthogonality residual $\|\mathbf{R}_m^\top \mathbf{R}_m - \mathbf{I}_d\|_F/\sqrt{d}$ is 7.96×10^{-7} for NIR and 7.54×10^{-7} for TIR, confirming that the implemented maps satisfy orthogonality to high numerical precision.

E.3.3 COVARIANCE STRUCTURE UNDER CANONICALIZATION

The preceding analysis describes the learned matrices. We now examine their effect on the feature distribution. Intuitively, the features of one modality form a cloud of points. Its covariance describes how widely the points spread along different directions. We show that canonicalization can change the position and orientation of this cloud while preserving its principal variances.

Consider the fixed affine transformation

$$\widetilde{\mathbf{z}} = \mathbf{R}_m^\top(\mathbf{z} - \boldsymbol{\mu}_m) + \boldsymbol{\mu}_{m_{\mathrm{ref}}}, \tag{80}$$

where $\boldsymbol{\mu}_m$ and $\boldsymbol{\mu}_{m_{\mathrm{ref}}}$ are the stored modality means. Let

$$\boldsymbol{\nu}_m = \mathbb{E}[\mathbf{z} \mid m] \tag{81}$$

denote the population feature mean, and let

$$\boldsymbol{\Sigma}_m = \mathbb{E}\big[(\mathbf{z} - \boldsymbol{\nu}_m)(\mathbf{z} - \boldsymbol{\nu}_m)^\top \mid m\big] \tag{82}$$

be the population covariance, assumed finite. We use $\boldsymbol{\nu}_m$ to distinguish the population mean from its running estimate $\boldsymbol{\mu}_m$.

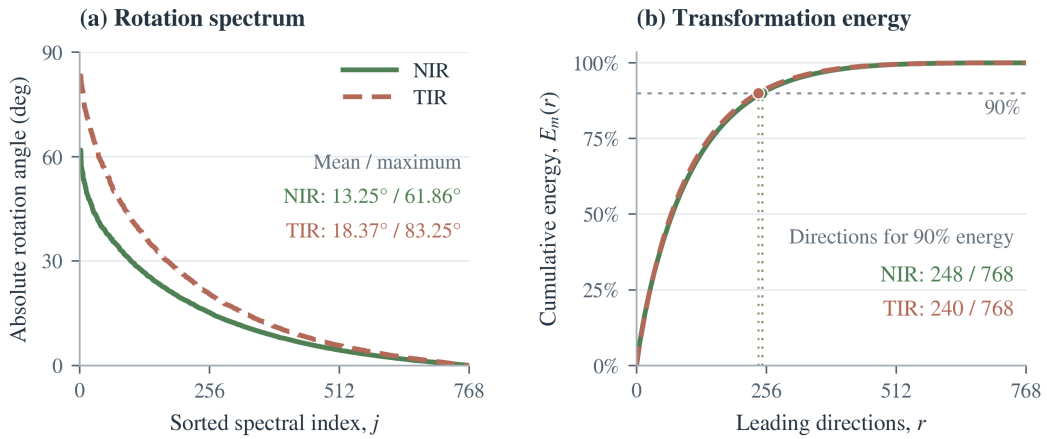

Figure 5: Analysis of the learned NIR and TIR rotations on WHU-MARS-1000. (a) Absolute rotation angles sorted in descending order; each rotation plane contributes two equal angles. (b) Cumulative transformation energy computed from $\mathbf{R}_m - \mathbf{I}_d$. The markers show that 248 directions for NIR and 240 for TIR account for 90% of the total energy. The RGB reference map is the identity and is omitted.

Preservation of principal variances. The transformed covariance satisfies

$$\widetilde{\mathbf{\Sigma}}_m = \mathbf{R}_m^\top \mathbf{\Sigma}_m \mathbf{R}_m, \qquad \lambda_j(\widetilde{\mathbf{\Sigma}}_m) = \lambda_j(\mathbf{\Sigma}_m). \tag{83}$$

The eigenvalues of a covariance matrix are the variances along its principal axes. This equality therefore states that these variances remain unchanged after canonicalization.

Proof. Taking the conditional expectation of Equation 80 gives

$$\widetilde{\boldsymbol{\nu}}_m = \mathbf{R}_m^\top (\boldsymbol{\nu}_m - \boldsymbol{\mu}_m) + \boldsymbol{\mu}_{m_{\text{ref}}}. \tag{84}$$

Subtracting this mean from the transformed feature cancels both translation terms:

$$\widetilde{\mathbf{z}} - \widetilde{\boldsymbol{\nu}}_m = \mathbf{R}_m^\top (\mathbf{z} - \boldsymbol{\nu}_m). \tag{85}$$

Substitution into the covariance definition yields

$$\begin{aligned} \widetilde{\mathbf{\Sigma}}_m &= \mathbb{E}\big[\mathbf{R}_m^\top (\mathbf{z} - \boldsymbol{\nu}_m)(\mathbf{z} - \boldsymbol{\nu}_m)^\top \mathbf{R}_m \mid m\big] \\ &= \mathbf{R}_m^\top \mathbf{\Sigma}_m \mathbf{R}_m. \end{aligned} \tag{86}$$

Write the eigendecomposition of the original covariance as

$$\mathbf{\Sigma}_m = \mathbf{U}_m \mathbf{\Lambda}_m \mathbf{U}_m^\top, \tag{87}$$

where $\mathbf{U}_m$ contains orthonormal eigenvectors and $\mathbf{\Lambda}_m$ contains the corresponding eigenvalues. Then

$$\widetilde{\mathbf{\Sigma}}_m = (\mathbf{R}_m^\top \mathbf{U}_m) \mathbf{\Lambda}_m (\mathbf{R}_m^\top \mathbf{U}_m)^\top. \tag{88}$$

The transformed eigenvectors remain orthonormal because

$$\begin{aligned} (\mathbf{R}_m^\top \mathbf{U}_m)^\top (\mathbf{R}_m^\top \mathbf{U}_m) &= \mathbf{U}_m^\top \mathbf{R}_m \mathbf{R}_m^\top \mathbf{U}_m \\ &= \mathbf{I}_d. \end{aligned} \tag{89}$$

Thus, the transformed covariance has the same eigenvalues, proving Equation 83.

Geometrically, canonicalization moves the feature cloud and rotates its principal axes without stretching or compressing it. For example, an elongated cloud remains equally elongated after the transformation, although its orientation can change. The total variance and covariance rank are also preserved. This explains how OCSFA adjusts modality coordinates while retaining the geometry within each modality before normalization.

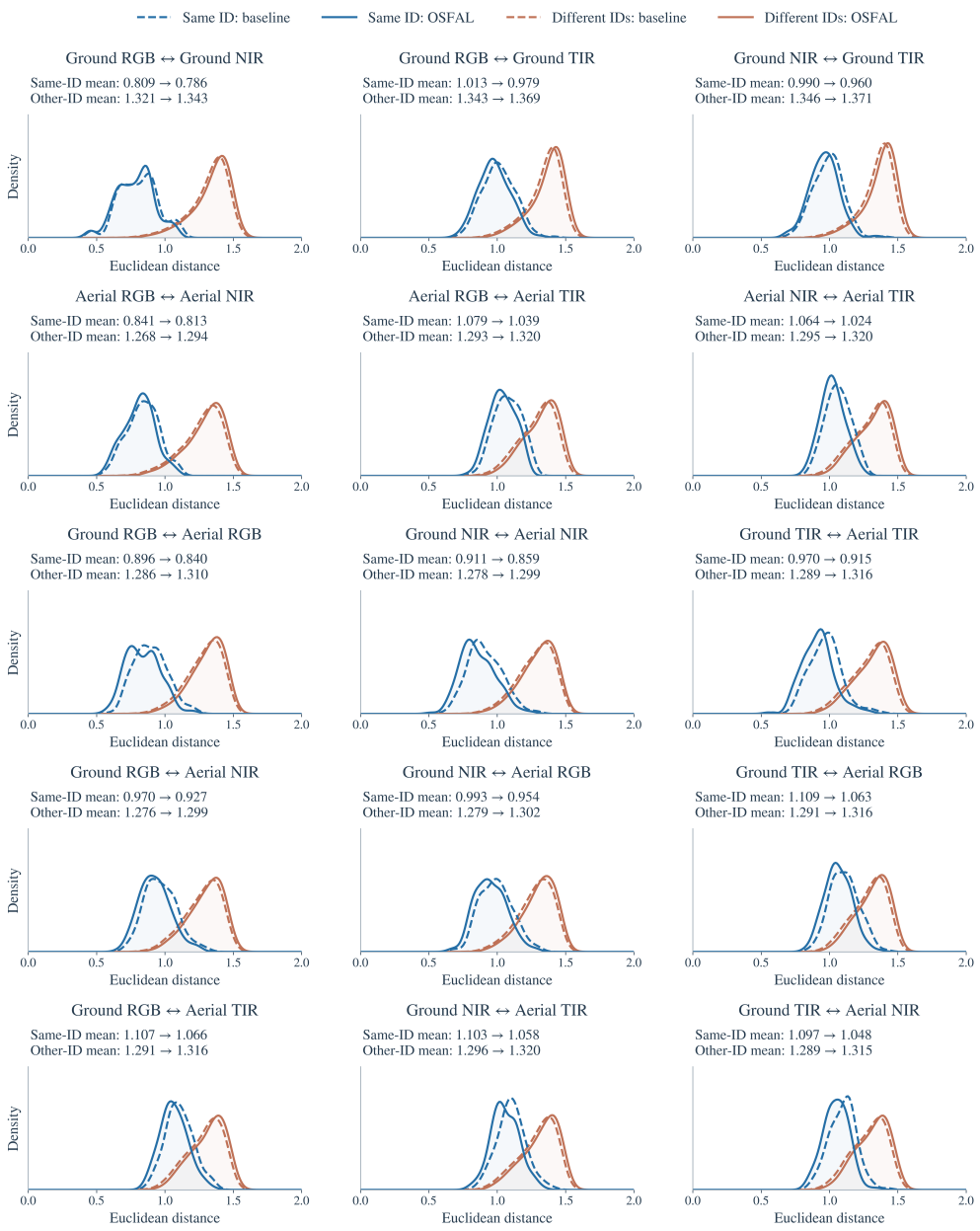

Figure 6: Euclidean-distance distributions of selected prototype pairs across modality–view conditions. Dashed and solid curves denote the baseline and OSFAL, respectively.

F Additional visulizations

F.1 Feature distance distribution

Figure 6 compares the Euclidean-distance distributions of the baseline and OSFAL across all 15 pairwise combinations of the six modality–view conditions. Each prototype is constructed by averaging the normalized features within an identity–modality–view cell, followed by ℓ_2 normalization of the resulting mean. Euclidean distances are then computed between prototypes from different modality–view conditions to characterize same-identity and different-identity relationships.

The blue distributions correspond to same-identity prototype pairs, whereas the orange distributions represent different-identity pairs. Dashed and solid curves denote the baseline and OSFAL,



Figure 7: Top-5 retrieval comparison on WHU-MARS-1000. Panels (a,b), (c,d), and (e,f) show selected improvements involving spectral changes, view changes, and both, respectively. Each case shows the baseline above OSFAL. Blue and red borders indicate correct and incorrect identities. G denotes Ground and A denotes Aerial.

respectively. Across all modality–view condition pairs, OSFAL consistently shifts the same-identity distributions toward smaller distances while moving the different-identity distributions toward larger distances. This simultaneous reduction in intra-identity distances and enlargement of inter-identity distances indicates that OSFAL produces a more compact and discriminative embedding space across heterogeneous modality and viewpoint conditions. For example, for the Ground RGB $\leftrightarrow$ Aerial RGB pair, the mean same-identity distance decreases from 0.896 to 0.840, whereas the mean different-identity distance increases from 1.286 to 1.310.

F.2 Qualitative retrieval illustration

Figure 7 compares the Top-5 retrieval results of the baseline and OSFAL on WHU-MARS-1000. Both models use the same queries and the complete mixed-scenario gallery, with same-camera images excluded. We present six improvement cases covering cross-spectral matching, cross-view matching, and matching across both factors. Blue and red borders indicate correct and incorrect identities, respectively; G and A denote Ground and Aerial views.

In panels (a) and (b), OSFAL retrieves correct cross-spectral matches at Rank-1, whereas the baseline ranks its first correct match at positions 5 and 4. Panels (c) and (d) illustrate cross-view improvements, with the first correct match moving from positions 183 and 19 to Rank-1. Panels (e) and (f) further show successful matches when both the spectral modality and view differ from the query. Across these six examples, all Top-5 results returned by OSFAL belong to the query identity. These

cases illustrate how the retrieval improvements manifest in individual rankings under heterogeneous sensing conditions.