

SARA: Structure-Aware Riemannian-Guided Alignment for Drone Image-Text Retrieval

Keyu Lu^{1,3}, Qing Ma^{2*}, Zhenyu Lu¹, and Cong Bai^{1,3}

¹ College of Computer Science and Technology, Zhejiang University of Technology,
Hangzhou 310023, China

² School of Mathematical Sciences, Zhejiang University of Technology, Hangzhou
310023, China

³ Zhejiang Key Laboratory of Visual Information Intelligent Processing, Hangzhou
310023, China

{202003160111,maqing,302023315230,congbai}@zjut.edu.cn

Abstract. Drone image-text retrieval (DITR) aims to associate aerial images with textual descriptions in complex low-altitude environments. However, intricate spatial layouts and long-range structural dependencies in drone imagery challenge conventional Euclidean embedding models, often leading to cross-modal misalignment. To address this issue, we propose SARA, a structure-aware Riemannian-guided alignment framework that incorporates Riemannian manifold modeling to complement semantic embeddings. Specifically, the Manifold Structural Feature Extraction (MSFE) module captures second-order structural dependencies by encoding local feature correlations as symmetric positive definite representations. The Log-Euclidean Structural Alignment (LESA) module performs geometry-consistent alignment in the tangent space, integrating semantic correspondence with structural coherence through distribution alignment and feature fusion. We further provide theoretical analysis demonstrating convergence and minimization of an upper bound on the expected alignment error. Experiments on ERA and UDV benchmarks demonstrate that SARA consistently improves retrieval performance over state-of-the-art methods, particularly in structurally complex or semantically ambiguous drone scenarios.

Keywords: Cross-Modal Retrieval · Riemannian Geometry · Structure Alignment

1 Introduction

With the rapid development of drone technologies, massive drone-captured images have been widely used in urban surveillance, disaster assessment, and environmental monitoring [4, 12]. Unlike ground-based or satellite imagery, drone images are typically captured from varying altitudes and oblique viewpoints, resulting in complex perspective variations and non-uniform spatial layouts of

* Corresponding author.

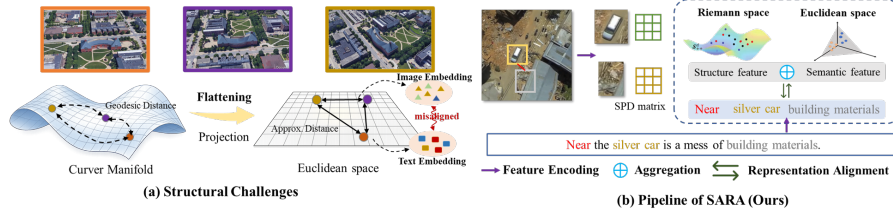


Fig. 1: (a) Challenges of drone image-text alignment: viewpoint variations induce nonlinear structural transformations that are difficult to capture under flat Euclidean embeddings. (b) Pipeline of the proposed SARA framework, which aggregates semantic and structural representations based on Riemannian manifold representations to achieve geometry-aware cross-modal alignment.

objects. As illustrated in Fig. 1(a), viewpoint changes induce complex spatial transformations in visual observations, altering relative object configurations while preserving underlying structural relationships. Such transformations are inherently nonlinear, and intrinsic geometric relations among objects cannot be faithfully captured by linear distances. When these correlated patterns are embedded into conventional Euclidean feature spaces, their nonlinear structures are approximated in flat geometry, potentially distorting structural dependencies. This mismatch between intrinsic visual geometry and the assumed Euclidean embedding space results in representation-level geometric inconsistency, which may degrade cross-modal alignment when image and text features are jointly learned in flat spaces [7, 17].

To mitigate these representation limitations, Riemannian geometry provides a complementary modeling perspective. A Riemannian manifold is equipped with a smoothly varying metric structure that enables intrinsic measurement of relational variations while accounting for local curvature and neighborhood consistency. In particular, the space of Symmetric Positive Definite (SPD) matrices offers a natural representation for modeling second-order feature correlations within a curved geometric domain [21]. Under metrics such as the Log-Euclidean metric, distances can be computed in a geometry-consistent manner that remains stable under common feature transformations [1, 9]. Consequently, Riemannian modeling provides a structural inductive bias that encourages curvature-aware and coordinate-consistent representation learning, complementing conventional Euclidean semantic embeddings in cross-modal representation tasks.

Building on this foundation, we propose the Structure-Aware Riemannian-Guided Alignment framework for drone image-text retrieval. As illustrated in Fig. 1(b), SARA leverages structural representations as guidance and jointly models them with semantic embeddings to promote geometry-consistent cross-modal alignment. It consists of two key modules that jointly model structural and semantic representations. The Manifold Structural Feature Extraction (MSFE) module encodes local visual regions as SPD matrices, enabling second-order aggregation that captures intrinsic spatial dependencies on the Riemannian mani-

fold. The Log-Euclidean Structural Alignment (LESA) module aligns manifold-based structural representations with semantic embeddings in the tangent space via a Log-Euclidean metric, ensuring curvature-consistent matching. Extensive experiments on two drone benchmarks demonstrate consistent improvements over state-of-the-art methods, validating the robustness and generality of geometry-aware alignment under complex aerial conditions.

Our main contributions are summarized as follows:

- We propose a novel retrieval framework, SARA (Structure-Aware Riemannian-Guided Alignment), which integrates Riemannian geometry with Euclidean semantic embeddings to achieve structure-aware cross-modal alignment in challenging UAV scenarios. By bridging semantic embedding and structural reasoning on the Riemannian manifold, SARA effectively alleviates distortion and misalignment issues caused by complex drone viewpoints.
- We design two complementary modules, MSFE and LESA, to jointly model semantic and structural representations. The Manifold Structural Feature Extraction encodes local visual features as SPD matrices to capture second-order structural dependencies, while the Log-Euclidean Structural Alignment performs geometry-consistent cross-modal alignment in the tangent space under the Log-Euclidean metric.
- We present a theoretical analysis showing that the proposed alignment loss converges and effectively minimizes an upper bound of the expected alignment error. Experiments on drone benchmarks (ERA and UDV) demonstrate consistent improvements, particularly in structurally complex scenes.

2 Related Work

Riemannian and Manifold Representation Learning. Riemannian geometry has been widely adopted for structure-preserving representation learning. Symmetric positive definite (SPD) manifolds model second-order feature statistics and have demonstrated strong capability in visual recognition tasks. SPDNet [26] introduces manifold-aware deep architectures to maintain geometric consistency during learning, and subsequent works extend SPD modeling to normalization, domain adaptation, and robust feature aggregation under Bures–Wasserstein or log-Euclidean metrics [11, 25]. Beyond SPD manifolds, hyperbolic embeddings [13] leverage negatively curved spaces to encode hierarchical semantics, highlighting the advantages of non-Euclidean geometry over Euclidean approximations. However, most existing manifold-based methods focus on unimodal representation learning and do not explicitly address cross-modal alignment under complex spatial variations.

Non-Euclidean Cross-Modal Retrieval. Recent cross-modal retrieval methods have begun to explore non-Euclidean embedding spaces for geometry-aware alignment. MERU [5] embeds image and text representations into hyperbolic space to capture hierarchical and asymmetric semantic relations, while HyCo-CLIP [19] introduces curvature-regularized contrastive learning to stabilize hyperbolic optimization and enhance representation consistency. These approaches

demonstrate the effectiveness of curved embedding spaces beyond conventional Euclidean similarity learning. Nevertheless, they primarily treat the manifold as an alternative global embedding geometry and focus on semantic-level alignment, without explicitly modeling structural feature correlations grounded in SPD statistics.

Drone Image–Text Retrieval. Drone image–text retrieval has attracted increasing attention due to flexible viewpoints and rich spatial details of low-altitude imagery. Existing methods typically rely on Euclidean embedding spaces and optimize cross-modal similarity through contrastive learning or matching objectives. Recent works incorporate contextual semantic reasoning [8], hierarchical cross-granularity alignment [23], transformer-based interaction modeling [27], and attention correction mechanisms [2, 29] to improve robustness under complex aerial scenes. However, Euclidean similarity modeling may struggle to preserve nonlinear spatial dependencies inherent in drone imagery. In contrast, our framework introduces SPD-based structural representations and performs alignment under Riemannian geometry, enabling geometry-consistent cross-modal similarity learning in aerial scenarios.

3 Proposed Method

Our Structure-Aware Riemannian-Guided Alignment (SARA) framework (Fig. 2) models the geometric structures of drone imagery and aligns them with textual descriptions in a geometry-preserving space. Section 3.1 details semantic feature extraction, Sec. 3.2 introduces the Manifold Structural Feature Extraction (MSFE) on the SPD manifold, and Sec. 3.3 presents the Log-Euclidean Structural Alignment (LESA) for tangent-space alignment. Finally, Sec. 3.4 provides a theoretical justification that proves minimizing the alignment loss reduces an upper bound of the expected alignment error, thereby ensuring convergence and stable optimization.

3.1 Semantic Feature Extraction

An input image–text pair (I, T) is encoded by modality-specific backbones to extract semantic representations. For the visual branch, the image $I \in \mathbb{R}^{3 \times H \times W}$ is fed into a vision encoder followed by a linear projection to obtain the semantic feature:

$$v_{\text{sem}} = W_v \text{Enc}_v(I) + b_v, \quad (1)$$

where $\text{Enc}_v(\cdot)$ denotes the vision encoder parameterized by θ_v , $W_v \in \mathbb{R}^{d \times d_v}$ and $b_v \in \mathbb{R}^d$ are learnable parameters.

Similarly, the text branch encodes the sentence $T = \{w_1, w_2, \dots, w_L\}$ through a text encoder and a linear projection as:

$$t_{\text{sem}} = W_t \text{Enc}_t(T) + b_t, \quad (2)$$

where $\text{Enc}_t(\cdot)$ is the text encoder parameterized by θ_t , $W_t \in \mathbb{R}^{d \times d_t}$ and $b_t \in \mathbb{R}^d$ are learnable parameters. The resulting embeddings v_{sem} and t_{sem} serve as global semantic representations for cross-modal alignment.

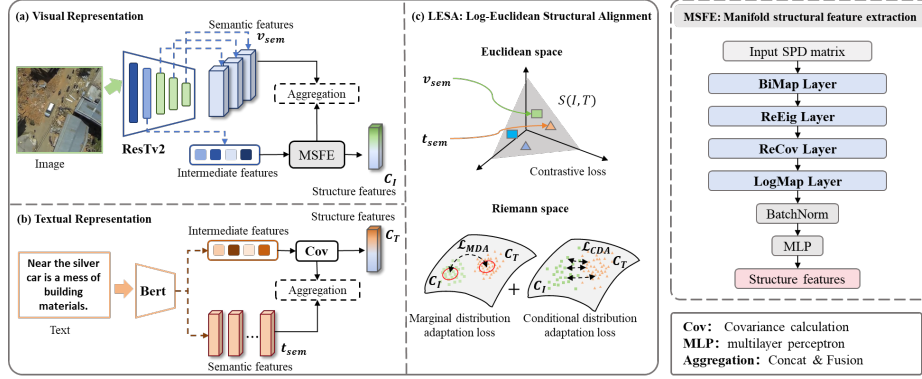


Fig. 2: Overview of the proposed SARA framework. It employs a semantic branch for content representation and a structural branch for manifold-based modeling on the SPD manifold. The MSFE module extracts second-order structural features, while the LESA module aligns image–text representations under the Log-Euclidean metric for geometry-aware cross-modal retrieval.

3.2 Manifold Structural Feature Extraction

While the semantic branch captures global meanings, drone imagery contains spatial regularities and geometric correlations that are not well represented in Euclidean space. The Manifold Structural Feature Extraction (MSFE) module models second-order statistics on a Riemannian manifold to obtain structure-aware representations for both modalities.

Visual Structural Representation In addition to the semantic feature v_{sem} , we utilize the intermediate representation from the fourth stage of the visual encoder $\text{Enc}_v(\cdot)$ to model structural correlations. The extracted feature map $\mathbf{F}_v = (\mathbf{f}_v^1, \dots, \mathbf{f}_v^N) \in \mathbb{R}^{C_v \times N}$, where C_v and N denote the channel dimension and the number of spatial tokens, is used to compute a second-order covariance descriptor as

$$\mathbf{C}_I^{(0)} = \frac{1}{N} \sum_{i=1}^N (\mathbf{f}_v^i - \bar{\mathbf{f}}_v)(\mathbf{f}_v^i - \bar{\mathbf{f}}_v)^\top + \epsilon \mathbf{I}, \quad (3)$$

where $\bar{\mathbf{f}}_v$ is the mean feature and $\epsilon \mathbf{I}$ ensures numerical stability. The resulting matrix $\mathbf{C}_I^{(0)} \in \text{Sym}_+^{C_v}$ is a Symmetric Positive Definite (SPD) representation residing on the Riemannian manifold, capturing inter-channel dependencies and spatial correlations that are overlooked by first-order embeddings.

To progressively refine $\mathbf{C}_I^{(0)}$, MSFE performs a hierarchical manifold-aware transformation consisting of the following layers:

- **BiMap Layer:** Performs an SPD-to-SPD linear projection to map the input matrix onto a lower-dimensional manifold while preserving positive definiteness:

$$\mathbf{C}^{(k)} = \mathbf{W}_k^\top \mathbf{C}^{(k-1)} \mathbf{W}_k, \quad k = 1, \dots, K, \quad (4)$$

where $\mathbf{W}_k \in \mathbb{R}^{d_{k-1} \times d_k}$ is semi-orthogonal on the Stiefel manifold $\mathcal{S}(d_{k-1}, d_k)$.

- **ReEig Layer:** Introduces non-linearity on the SPD manifold by rectifying the eigenvalues of the input matrix:

$$\tilde{\mathbf{C}}^{(k)} = \mathbf{U} \max(\mathbf{\Sigma}, \zeta \mathbf{I}) \mathbf{U}^\top, \quad (5)$$

where $\mathbf{C}^{(k)} = \mathbf{U} \mathbf{\Sigma} \mathbf{U}^\top$ is the eigendecomposition of the input SPD matrix, $\zeta > 0$ is a rectification threshold ensuring positive definiteness, and $\mathbf{I}$ is the identity matrix.

- **ReCov Layer:** Enhances off-diagonal elements to strengthen inter-channel correlations:

$$\bar{\mathbf{C}}^{(k)} = \tilde{\mathbf{C}}^{(k)} + \lambda(\tilde{\mathbf{C}}^{(k)} - \text{diag}(\tilde{\mathbf{C}}^{(k)})), \quad (6)$$

where $\text{diag}(\cdot)$ denotes the diagonal operator that preserves only the diagonal entries of a matrix, and $\lambda > 0$ controls the enhancement strength.

- **LogEig Layer:** Maps the SPD matrix from the manifold $\text{Sym}_+^{d_k}$ to its tangent space using the Log-Euclidean metric:

$$\hat{\mathbf{C}}^{(k)} = \log(\bar{\mathbf{C}}^{(k)}) = \mathbf{U} \log(\mathbf{\Sigma}) \mathbf{U}^\top, \quad (7)$$

where $\bar{\mathbf{C}}^{(k)}$ is decomposed via eigendecomposition. This linearizes the manifold for Euclidean computations while preserving intrinsic geometry.

The final structure-aware visual descriptor is $\mathbf{C}_I = \hat{\mathbf{C}}^{(K)} \in \mathbb{R}^{d_K \times d_K}$ after K SPD transformations.

Textual Structural Representation Analogous to the visual branch, we construct a structural representation for text. Given token-level embeddings from the last transformer layer $\text{Enc}_t(\cdot)$, we obtain $\mathbf{F}_t = (\mathbf{f}_t^1, \dots, \mathbf{f}_t^L) \in \mathbb{R}^{C_t \times L}$, where L is the token number and C_t the hidden dimension. The features are refined by a lightweight MLP and converted into a covariance descriptor:

$$\mathbf{C} = \text{Cov}(\text{MLP}(\mathbf{F}_t)), \quad \mathbf{C}_T = \log(\mathbf{W}_s \mathbf{C} \mathbf{W}_s^\top), \quad (8)$$

where $\mathbf{W}_s$ is a shared semi-orthogonal mapping ensuring geometric consistency with the visual SPD representation $\mathbf{C}_I$. The resulting $\mathbf{C}_T \in \text{Sym}_+^C$ encodes inter-token dependencies, serving as the textual counterpart to $\mathbf{C}_I$ for manifold-based alignment.

3.3 Log-Euclidean Structural Alignment

The Log-Euclidean Structural Alignment (LESA) module aligns modalities by leveraging structural representations on the Riemannian manifold and semantic embeddings in Euclidean space. It performs intra-modal alignment of SPD features in the tangent space and semantic-structural fusion for unified retrieval embedding.

Intra-Modal Riemannian Alignment Given manifold-aware features from the MSFE module, LESA first aligns image and text structures in a geometry-consistent space. Let $\mathbf{C}_I^{(i)}$ and $\mathbf{C}_T^{(i)}$ denote the i -th image and text structural matrices in a batch of size B :

$$\mathcal{I} = \{\mathbf{C}_I^{(i)}\}_{i=1}^B, \quad \mathcal{T} = \{\mathbf{C}_T^{(i)}\}_{i=1}^B.$$

Marginal Distribution Alignment (MDA) Although the Log-Euclidean mapping transforms SPD representations into a Euclidean space, the visual and textual structural features may still follow different global distributions due to modality-specific biases. Such discrepancy can hinder cross-modal consistency, even if local structures are well aligned. To mitigate this issue, we perform marginal distribution alignment by encouraging the Fréchet means of the two modalities to coincide:

$$\mathcal{L}_{\text{MDA}} = \left\| \frac{1}{B} \sum_{i=1}^B \mathbf{C}_I^{(i)} - \frac{1}{B} \sum_{i=1}^B \mathbf{C}_T^{(i)} \right\|_F^2. \quad (9)$$

By aligning their first-order statistics, MDA reduces the global distribution gap between modalities, ensuring that both structural representations lie in a shared manifold-centered region and facilitating more consistent cross-modal matching.

Conditional Distribution Alignment (CDA) While MDA aligns the global statistics of visual and textual structures, it ignores the fine-grained relationships between semantically matched pairs. In practice, samples with the same semantics should also exhibit similar structural representations. To enforce such local consistency, we perform conditional distribution alignment by minimizing the discrepancy between paired samples:

$$\mathcal{L}_{\text{CDA}} = \frac{1}{N_{\text{pos}}} \sum_{i=1}^{N_{\text{pos}}} \left\| \mathbf{C}_I^{(i)} - \mathbf{C}_T^{(i)} \right\|_F^2, \quad (10)$$

where N_{pos} denotes the number of positive image–text pairs. By aligning structural features under semantic correspondence, CDA narrows the conditional gap between modalities and enhances fine-grained consistency beyond the global distribution alignment achieved by MDA.

Semantic Supervision To preserve discriminative semantics, we apply bidirectional contrastive losses:

$$\mathcal{L}_{\text{v2t}} = -\frac{1}{B} \sum_{i=1}^B \log \frac{\exp(s_{i,i}/\tau)}{\sum_{j=1}^B \exp(s_{i,j}/\tau)}, \quad (11)$$

$$\mathcal{L}_{\text{t2v}} = -\frac{1}{B} \sum_{i=1}^B \log \frac{\exp(s_{i,i}/\tau)}{\sum_{j=1}^B \exp(s_{j,i}/\tau)}. \quad (12)$$

where $s_{i,j} = v_{\text{sem}}^{(i)} \top t_{\text{sem}}^{(j)}$. The total semantic loss is

$$\mathcal{L}_{\text{sem}} = \frac{1}{2}(\mathcal{L}_{\text{v2t}} + \mathcal{L}_{\text{t2v}}). \quad (13)$$

Semantic-Structural Fusion and Second-Stage Alignment Although manifold alignment reduces the cross-modal gap, semantic and structural features remain in separate spaces. To leverage their complementarity, a learnable fusion module integrates them into unified representations:

$$\mathbf{z}_I = \mathbf{W}_f[v_{\text{sem}} \parallel \text{vec}(\mathbf{C}_I)], \quad \mathbf{z}_T = \mathbf{W}_f[t_{\text{sem}} \parallel \text{vec}(\mathbf{C}_T)], \quad (14)$$

where $\parallel$ denotes vector concatenation and $\text{vec}(\cdot)$ vectorizes an SPD matrix. The fused embeddings are normalized, refined by a lightweight MLP, and aligned via a symmetric contrastive loss:

$$\mathcal{L}_{v2t}^* = -\frac{1}{B} \sum_{i=1}^B \log \frac{\exp(\text{sim}(\mathbf{z}_I^{(i)}, \mathbf{z}_T^{(i)})/\tau)}{\sum_{j=1}^B \exp(\text{sim}(\mathbf{z}_I^{(i)}, \mathbf{z}_T^{(j)})/\tau)}, \quad (15)$$

$$\mathcal{L}_{t2v}^* = -\frac{1}{B} \sum_{i=1}^B \log \frac{\exp(\text{sim}(\mathbf{z}_T^{(i)}, \mathbf{z}_I^{(i)})/\tau)}{\sum_{j=1}^B \exp(\text{sim}(\mathbf{z}_T^{(i)}, \mathbf{z}_I^{(j)})/\tau)}, \quad (16)$$

$$\mathcal{L}_{\text{fusion}} = \frac{1}{2}(\mathcal{L}_{v2t}^* + \mathcal{L}_{t2v}^*). \quad (17)$$

Overall Objective The final training loss integrates all components:

$$\mathcal{L}_{\text{SARA}} = \mathcal{L}_{\text{sem}} + \lambda_m \mathcal{L}_{\text{MDA}} + \lambda_c \mathcal{L}_{\text{CDA}} + \alpha \mathcal{L}_{\text{fusion}}, \quad (18)$$

where λ_m , λ_c , and α balance global, local, and fusion alignments. This hierarchical optimization progressively aligns modalities from the Riemannian manifold to a unified Euclidean space.

3.4 Theoretical Analysis

We provide a theoretical justification for the proposed Riemannian alignment objective by showing that minimizing the expected geometric discrepancy between image-text pairs directly controls the expected matching error.

Theorem 1 (Upper Bound of Expected Alignment Error). *Let the similarity function $s(\cdot)$ be L -Lipschitz continuous with respect to the combined distance $d_{\text{comb}}(v, t)$, and define the training loss as*

$$\mathcal{L}_{\text{SARA}} = \mathbb{E}_{(v,t) \sim \mathcal{D}}[d_{\text{comb}}(v, t)]. \quad (19)$$

which is inherently a function of the paired variables (v, t) .

Let $\phi_\gamma(z) = \max(0, 1 - z/\gamma)$ denote the hinge-type surrogate loss with margin $\gamma > 0$. Then, the expected alignment error satisfies

$$\mathcal{E} = \mathbb{E}_{(v,t) \sim \mathcal{D}}[\mathbb{I}\{\text{mismatch}(v, t)\}] \leq \frac{L}{\gamma} \mathcal{L}_{\text{SARA}}. \quad (20)$$

Proof. The proof proceeds as follows.

1. The indicator loss is upper-bounded by the hinge-type surrogate:

$$\mathbb{I}\{\text{mismatch}(v, t)\} \leq \phi_\gamma(y \cdot s(v, t)), \quad (21)$$

where $y \in \{+1, -1\}$ denotes whether the pair (v, t) is matched.

2. For the ideal similarity function s^* that perfectly separates matched ($y = +1$) and unmatched ($y = -1$) pairs with margin γ , we have

$$y \cdot s^* \geq \gamma. \quad (22)$$

3. The L -Lipschitz property of the similarity function implies

$$|s(v, t) - s^*(v, t)| \leq L d_{\text{comb}}(v, t). \quad (23)$$

4. Combining these results, we can bound the surrogate loss:

$$\begin{aligned} \phi_\gamma(y \cdot s(v, t)) &= \max\left(0, 1 - \frac{y s(v, t)}{\gamma}\right) \\ &= \max\left(0, 1 - \frac{y s^*}{\gamma} + \frac{y(s^* - s(v, t))}{\gamma}\right) \\ &\leq \max\left(0, 1 - \frac{y s^*}{\gamma}\right) + \frac{|s^* - s(v, t)|}{\gamma} \\ &\leq \max\left(0, 1 - \frac{\gamma}{\gamma}\right) + \frac{L d_{\text{comb}}(v, t)}{\gamma} \\ &\leq \frac{L}{\gamma} d_{\text{comb}}(v, t). \end{aligned} \quad (24)$$

The first inequality follows from the triangle inequality, and the second from $y \cdot s^* \geq \gamma$ and the Lipschitz condition.

5. Taking expectations over $\mathcal{D}$ yields:

$$\mathbb{E}[\phi_\gamma(y \cdot s(v, t))] \leq \frac{L}{\gamma} \mathbb{E}_{(v, t) \sim \mathcal{D}}[d_{\text{comb}}(v, t)] = \frac{L}{\gamma} \mathcal{L}_{\text{SARA}}. \quad (25)$$

6. Combining with step 1 completes the proof:

$$\mathcal{E} \leq \frac{L}{\gamma} \mathcal{L}_{\text{SARA}}. \quad (26)$$

4 Experiments

4.1 Datasets and Implementation Details

Datasets. We evaluate SARA on two drone image–text retrieval (DIT) benchmarks, namely the ERA DIT dataset and the UDV DIT dataset [8], both designed for low-altitude drone scenarios with diverse viewpoints and complex

Table 1: Comparison results on ERA and UDV DIT datasets. Methods marked with * operate in non-Euclidean manifold spaces.

Method	Category	Source	ERA dataset							UDV dataset						
			Text Retrieval			Image Retrieval			mR	Text Retrieval			Image Retrieval			mR
			R@1	R@5	R@10	R@1	R@5	R@10		R@1	R@5	R@10	R@1	R@5	R@10	
VSE++	CNN-RNN	BMVC18	9.79	30.40	42.90	10.13	35.20	53.91	30.39	1.76	7.21	12.98	2.27	8.68	15.04	8.01
PVSE K=2	CNN-RNN	CVPR19	11.31	32.60	46.95	11.04	35.57	51.65	31.52	3.16	12.53	20.81	3.35	12.84	21.02	12.29
PVSE K=1	CNN-RNN	CVPR19	9.96	33.95	47.97	11.14	36.08	53.75	32.14	2.90	12.55	21.55	3.23	12.62	20.94	12.30
CLIP	VLP	ICML21	11.31	31.92	43.91	12.73	37.33	51.52	31.45	6.35	20.26	29.99	4.25	16.15	26.09	17.18
PCME	CNN-RNN	CVPR21	14.69	35.30	49.15	13.85	42.87	60.64	36.08	6.51	21.83	31.92	6.37	21.34	32.76	19.18
AMFMN-soft	CNN-RNN	TGRS22	14.35	38.01	52.02	14.18	46.79	62.87	38.04	6.18	20.86	30.70	5.98	21.33	32.47	19.59
AMFMN-sim	CNN-RNN	TGRS22	14.02	34.12	51.52	13.75	43.41	59.59	36.06	6.56	21.40	31.08	6.17	20.93	32.08	19.70
AMFMN-fusion	CNN-RNN	TGRS22	15.20	36.99	50.33	11.62	42.26	60.51	36.15	6.23	20.88	31.11	6.32	20.84	31.46	19.47
GaLR	CNN-RNN	TGRS22	12.38	36.59	50.90	14.03	45.15	64.54	37.27	6.11	20.01	30.06	5.52	19.47	30.85	18.67
MSA	Transformer	TGRS24	18.04	50.54	64.86	19.93	41.22	55.74	41.72	4.75	18.26	29.62	4.85	16.37	24.88	16.45
VCSR	CNN-RNN	TGRS24	15.65	38.28	53.49	13.69	46.31	66.37	38.96	7.33	22.25	32.11	6.45	22.25	33.61	20.67
HCCM	Transformer	ACM MM25	19.93	39.19	56.76	18.58	45.20	59.93	39.93	7.97	19.60	28.38	7.43	22.60	29.97	19.33
MERU*	Transformer	TGRS24	6.22	21.37	32.98	7.41	21.06	32.33	20.18	2.63	8.92	15.23	2.38	13.51	15.24	9.12
HyCoCLIP*	VLP	ICLR25	7.54	20.57	33.21	8.11	20.85	34.56	22.31	3.12	10.45	17.02	3.38	11.02	17.55	10.42
SARA	Transformer		23.98	48.64	63.17	21.55	56.82	76.01	48.36	8.09	24.55	34.68	6.88	23.41	35.35	22.16

backgrounds. The ERA DIT dataset contains 2,864 images, each paired with five human-annotated captions. The UDV DIT dataset includes 10,693 image-text pairs collected from VisDrone, University-1652, and additional drone footage. Compared to ERA, UDV covers broader urban and rural scenes with higher structural and semantic diversity. We follow the standard data splits (70%/20%/10% for ERA and 60%/20%/20% for UDV) for training, validation, and testing.

Metrics. We evaluate two retrieval tasks: Image-to-Text (I2T) and Text-to-Image (T2I). Retrieval performance is measured using Recall@K (R@1, R@5, R@10) and mean Recall (mR), which jointly reflect bidirectional cross-modal alignment performance.

Implementation Details. We adopt RestV2-Base (pre-trained on ImageNet-1K) as the vision encoder and BERT-base-uncased as the text encoder. Both encoders are projected to a shared 512-dimensional embedding space for multi-modal alignment. All models are fine-tuned end-to-end using Adam or Riemannian Adam optimizer with an initial learning rate of 2×10^{-4} . The temperature parameter in the contrastive loss is set to 0.07. Training is conducted for 50 epochs on a single NVIDIA RTX A6000 GPU, with batch sizes of 120 for ERA and 210 for UDV.

4.2 Performance Comparisons

We compare SARA with state-of-the-art methods on ERA and UDV for both I2T and T2I retrieval. All results are averaged over three runs.

State-of-the-art Methods We compare our SARA framework with several state-of-the-art cross-modal retrieval methods across different architectural categories (Tab. 1), including CNN-RNN models that combine convolutional visual encoders with recurrent text encoders (VSE++ [6], PVSE [24], PCME [3], AMFMN [30], GaLR [31], VCSR [8]), Transformer-based methods using self-attention for cross-modal interaction (MSA [28], HCCM [23]), vision-language pre-training (VLP) models trained on large-scale image-text data (CLIP [22]),

and manifold-based approaches that learn representations in non-Euclidean spaces (MERU [5], HyCoCLIP [19]).

Comparison with Euclidean Methods. As shown in Tab. 1, SARA consistently outperforms Euclidean embedding methods on both ERA and UDV datasets. On ERA, compared with the strongest Euclidean baseline HCCM, SARA achieves an mR of 48.36, yielding a substantial +8.43 absolute improvement. On UDV, SARA further improves mR from 19.33 (HCCM) to 22.16. Notably, SARA also surpasses all other Euclidean approaches, including VCSR (38.96 / 20.67) and MSA (41.72 / 16.45), demonstrating consistent advantages in modeling cross-modal structural alignment under complex aerial scenarios.

Comparison with Manifold-based Methods. We compare SARA with non-Euclidean manifold approaches, including MERU and HyCoCLIP. Although these methods introduce hyperbolic or Riemannian embeddings, their performance remains lower (*e.g.* MERU achieves 20.18 mR on ERA and 9.12 mR on UDV). Similarly, HyCoCLIP obtains 22.31 mR on ERA and 10.42 mR on UDV. These results indicate that directly adopting manifold geometry is insufficient for drone image-text retrieval. In contrast, SARA integrates structure-aware SPD representations with a Riemannian alignment objective, enabling effective coupling between geometric structure and semantic similarity.

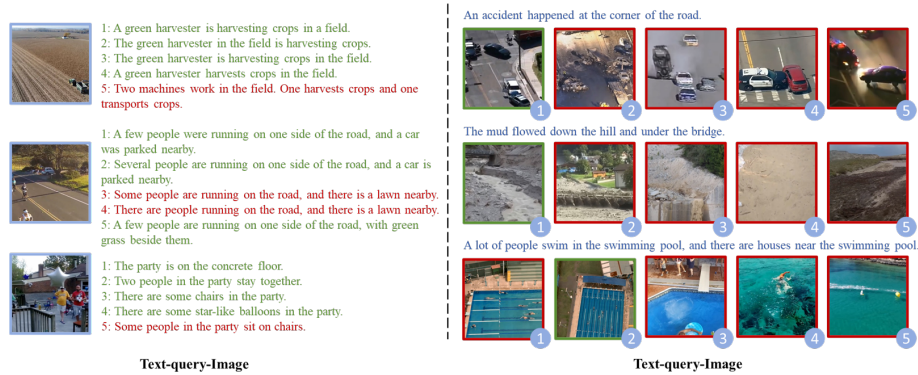


Fig. 3: Visualization of the top-5 retrieval results for Image-to-Text (left) and Text-to-Image (right) tasks, where green indicates correct matches and red denotes mismatches.

Visualization of results Fig. 3 illustrates the top-5 retrieval results for both Image-to-Text (left) and Text-to-Image (right) tasks. In Image-to-Text retrieval, most of the correctly matched captions (in green) accurately describe the visual content, showing that the model can effectively capture the semantic correspondence between image and text. Similarly, in Text-to-Image retrieval, the top-ranked retrieved images (in green boxes) are visually consistent with the textual descriptions. However, a few mismatched samples (in red) remain semantically similar to the query, indicating that the model may still struggle with fine-

Table 2: Ablation study of SARA components on ERA DIT and UDV DIT datasets.

Method	MSFE		LESA		ERA dataset						mR	UDV dataset						mR
	VSF	TSF	MDA	CDA	Image→Text			Text→Image				Image→Text			Text→Image			
					R@1	R@5	R@10	R@1	R@5	R@10		R@1	R@5	R@10	R@1	R@5	R@10	
Baseline	×	×	×	×	17.90	53.31	70.81	18.91	45.60	58.10	44.11	6.85	22.50	32.44	6.55	22.95	34.84	21.02
m_1	✓	×	×	×	20.06	52.77	72.22	19.93	45.27	60.47	45.12	7.21	23.10	33.15	6.67	23.04	34.91	21.35
m_2	×	✓	×	×	12.63	47.31	63.37	15.25	36.28	54.48	36.86	6.30	21.88	31.12	6.01	22.41	33.85	20.26
m_3	✓	✓	×	×	18.71	55.00	70.20	22.63	46.28	59.79	45.43	7.38	23.55	33.44	6.79	23.21	35.05	21.57
m_4	×	×	✓	✓	19.02	54.85	72.31	19.76	46.88	60.21	45.50	7.12	23.16	33.28	6.74	23.26	35.11	21.44
m_5	✓	✓	✓	×	19.45	55.20	73.04	22.63	45.94	62.83	46.62	7.66	23.71	33.92	6.81	23.26	35.10	21.79
m_6	✓	✓	×	✓	21.09	53.24	71.41	22.29	48.31	63.51	46.81	7.75	24.12	34.10	6.83	23.33	35.18	21.89
SARA	✓	✓	✓	✓	21.55	56.82	76.01	23.98	48.64	63.17	48.36	8.09	24.55	34.68	6.88	23.41	35.35	22.16

grained semantic discrimination or visually confusing scenarios. Overall, these results demonstrate strong cross-modal retrieval capability while suggesting potential improvements in distinguishing subtle contextual differences.

4.3 Ablation Studies

We conduct ablation experiments on both ERA DIT and UDV datasets to assess the contribution of each SARA component, using mean recall (mR) as the primary evaluation metric (Tab. 2).

Effect of MSFE MSFE significantly improves retrieval performance over the semantic-only baseline on both datasets. On ERA, adding Visual Structural Features (VSF) from C_I increases mR from 44.11 to 45.12 (I2T R@1 20.06; T2I R@1 19.93), while on UDV mR improves from 21.02 to 21.35 (I2T R@1 7.21). This confirms that modeling visual second-order correlations enhances geometry-aware alignment. In contrast, using only Textual Structural Features (TSF) from C_T degrades performance (ERA: 36.86 mR; UDV: 20.26 mR). This suggests that textual token correlations, which are largely syntactic or co-occurrence driven, do not reliably encode spatial structure; imposing SPD modeling on such signals may amplify spurious dependencies and hinder cross-modal consistency. Combining VSF and TSF yields the best performance (45.43 mR on ERA; 21.57 on UDV), indicating that structural modeling is most effective when grounded in visual geometry and complemented by textual relational cues.

Effect of LESA The Local-Enhanced Semantic Alignment (LESA) module further boosts performance through distribution-level alignment on the SPD manifold. Adding marginal distribution alignment (MDA) increases mR from 45.43 to 46.62 on ERA and from 21.57 to 21.79 on UDV, mitigating global distribution bias between modalities. Conditional distribution alignment (CDA) further improves mR to 46.81 (ERA) and 21.89 (UDV), highlighting the importance of pairwise alignment for instance-level matching. Integrating both MDA and CDA in SARA achieves the best results—mR 48.36 on ERA and 22.16 on UDV—with I2T R@1 gains of 3.84% and 1.24%, respectively. To isolate loss effects, we tested MDA/CDA only in the semantic space, where they consistently underperformed SARA, confirming that the gains stem from combining MDA/CDA with SPD-based structural representations. These results demonstrate that MDA and CDA

are complementary: global consistency and local pairwise alignment together enable robust Riemannian alignment on the SPD manifold, substantially enhancing cross-modal correspondence in both datasets.

4.4 Parameter Sensitivity

We evaluate the impact of key hyperparameters in LESA: the MDA and CDA weights (λ_m , λ_c) and the structural loss weight α , keeping other settings fixed.

Effect of λ_m and λ_c To examine the impact of MDA and CDA, we vary λ_m and λ_c over 0.01, 0.1, 0.3, 0.7, 1.0, 2.0 and report mR in Fig. 4. Results reveal a symmetric trend around moderate values, with peak performance (48.36) at $\lambda_m = 0.3$ and $\lambda_c = 0.7$, highlighting the importance of balanced regularization. Insufficient weights lead to alignment inconsistency, while values exceeding 1.0 over-constrain the model, impairing semantic discrimination and reducing retrieval accuracy.

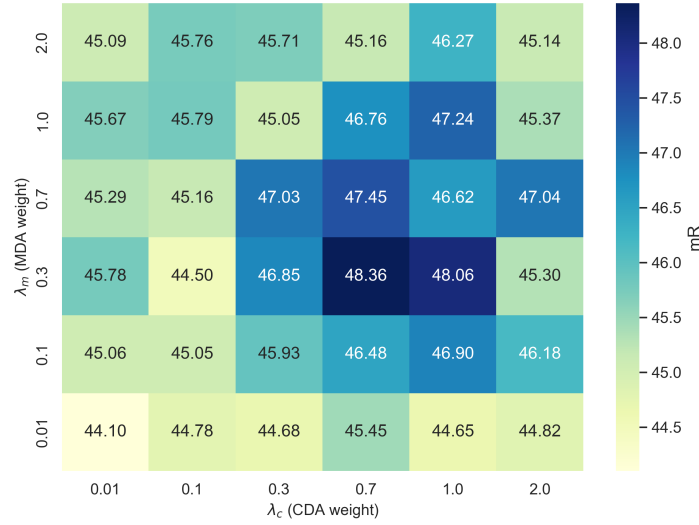
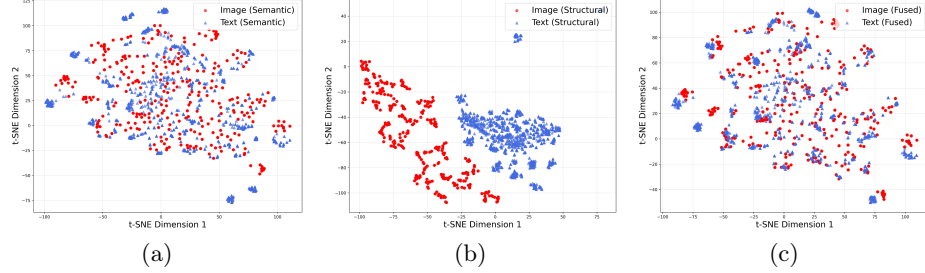


Fig. 4: Sensitivity of LESA module loss weights λ_m and λ_c on the ERA DIT Dataset.

Effect of α . As shown in Tab. 3, we vary the structural loss weight $\alpha \in \{0.01, 0.1, 0.3, 0.5, 0.7, 1.0\}$. On both the ERA and RSITMD datasets, the best performance is achieved at $\alpha = 0.7$, indicating that moderate structural regularization enhances semantic–structural complementarity, whereas an excessively large weight weakens cross-modal alignment.

Table 3: Effect of Structural Loss Weight (α) on ERA and RSITMD Datasets

Structural Loss α	ERA dataset							RSITMD						
	Text Retrieval			Image Retrieval			mR	Text Retrieval			Image Retrieval			mR
	R@1	R@5	R@10	R@1	R@5	R@10		R@1	R@5	R@10	R@1	R@5	R@10	
$\alpha = 0.01$	19.84	45.47	60.22	20.17	52.82	73.21	45.23	19.69	38.47	51.93	18.28	44.33	60.83	38.92
$\alpha = 0.1$	21.62	48.31	61.48	20.40	55.33	73.58	46.79	20.32	39.29	52.17	18.34	44.85	61.90	39.48
$\alpha = 0.3$	21.95	43.91	58.44	18.91	55.40	75.47	45.68	20.41	39.15	52.48	18.94	44.87	61.32	39.53
$\alpha = 0.5$	21.62	46.28	61.48	21.01	54.79	72.70	46.31	21.03	39.86	53.02	19.42	46.73	62.81	40.20
$\alpha = 0.7$	23.98	48.64	63.17	21.55	56.82	76.01	48.36	21.58	40.28	53.55	19.75	46.28	62.65	40.68
$\alpha = 1.0$	20.60	43.24	59.12	17.02	51.41	69.72	43.52	19.85	38.64	52.93	18.37	44.92	60.74	39.21

**Fig. 5:** t-SNE visualizations of (a) the semantic feature space, (b) the structural feature space, and (c) the fused feature space by SARA. Red squares denote image embeddings, and blue triangles denote text embeddings.

4.5 Embedding Visualization

To better understand SARA’s representation capability, we visualize semantic, structural, and fused feature spaces using t-SNE in Fig. 5. The semantic space in (a) captures contextual similarity but lacks geometric coherence, resulting in partial alignment. The structural space in (b) encodes spatial dependencies and geometric relations, preserving scene topology and reducing viewpoint distortion. While structural features alone cannot achieve full cross-modal alignment, they provide essential geometric cues for robust retrieval. Finally, the fused space in (c), produced by SARA’s integration of both semantic and structural features on the Riemannian manifold, forms a compact and well-aligned distribution in which image and text samples are closely interleaved.

4.6 Universal Validity of SARA on Remote Sensing Datasets

To further verify its universality, we evaluate SARA on two standard remote sensing image-text retrieval datasets, RSICD [15] and RSITMD [30]. We compare SARA with five representative methods, PIR [20], MaPLE [10], UniAdapter [14], DOVE [16], and SIPN [18]. As shown in Tab. 4, SARA achieves mR scores of 29.18 on RSICD and 40.68 on RSITMD, outperforming existing methods in overall retrieval performance. Notably, remote sensing images typically exhibit strong spatial structure and complex object layouts. By modeling second-order structural statistics as SPD matrices and performing Riemannian-guided alignment, SARA effectively captures long-range spatial dependencies and geometry-aware

Table 4: Retrieval performance of SARA on remote sensing datasets.

Method	Category	Source	RSICD							RSITMD						
			Text Retrieval			Image Retrieval			mR	Text Retrieval			Image Retrieval			mR
PIR	Transformer	ACM MM23	8.61	23.57	35.40	5.49	21.84	38.04	21.99	11.14	33.14	45.21	8.60	33.02	56.43	30.99
MaPLE	VLP	CVPR23	7.93	26.78	42.42	6.88	24.47	40.12	24.77	16.46	39.97	53.02	13.59	43.25	61.92	38.12
UniAdapter	VLP	ICLR24	12.65	28.81	42.74	9.61	30.06	46.16	27.54	19.86	36.32	51.28	17.54	44.89	65.46	39.23
DOVE	CNN-RNN	TGRS24	10.54	25.31	38.27	7.93	23.54	40.13	23.79	13.51	35.25	48.12	10.21	34.52	58.45	33.29
SIPN	Transformer	TGRS25	13.81	28.92	44.83	9.72	30.66	46.21	28.58	18.41	45.82	58.46	16.04	43.17	58.97	38.93
SARA	Transformer		10.83	29.58	47.22	12.42	31.33	46.68	29.18	19.75	46.28	62.65	21.58	40.28	53.55	40.68

correlations that are difficult to model in Euclidean space. This makes SARA particularly well-suited for structurally rich remote sensing scenarios.

5 Conclusion

In this paper, we proposed SARA, a Structure-Aware Riemannian-Guided Alignment framework for drone image-text retrieval. By representing features as symmetric positive definite (SPD) matrices and leveraging the Log-Euclidean metric on a Riemannian manifold, SARA captures fine-grained structural correlations neglected by conventional Euclidean methods. Theoretical analysis confirms that the proposed objective minimizes the expected alignment error, while extensive experiments on drone benchmarks demonstrate that SARA consistently outperforms state-of-the-art methods, particularly in complex scenes. Future work will extend this Riemannian paradigm to other aerial vision tasks and explore learnable Riemannian metrics.

Acknowledgements

This work is partially supported by Natural Science Foundation of China under Grant No.U20A20196, and the Zhejiang Provincial Natural Science Foundation of China under Grant No.LRG25F020002 and No.LY24F020020.

References

1. Arsigny, V., Fillard, P., Pennec, X., Ayache, N.: Geometric means in a novel vector space structure on symmetric positive-definite matrices. *SIAM Journal on Matrix Analysis and Applications* **29**(1), 328–347 (2007)
2. Chu, M., Zheng, Z., Ji, W., Wang, T., Chua, T.S.: Towards natural language-guided drones: GeoText-1652 benchmark with spatial relation matching. In: *Proceedings of the European Conference on Computer Vision*. pp. 213–231. Springer (2024)
3. Chun, S., Oh, S.J., De Rezende, R.S., Kalantidis, Y., Larlus, D.: Probabilistic embeddings for cross-modal retrieval. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 8415–8424 (2021)
4. Colomina, I., Molina, P.: Unmanned aerial systems for photogrammetry and remote sensing: A review. *ISPRS Journal of Photogrammetry and Remote Sensing* **92**, 79–97 (2014)

5. Desai, K., Nickel, M., Rajpurohit, T., Johnson, J., Vedantam, S.R.: Hyperbolic image-text representations. In: International Conference on Machine Learning. pp. 7694–7731. PMLR (2023)
6. Faghri, F., Fleet, D.J., Kiros, J.R., Fidler, S.: Vse++: Improving visual-semantic embeddings with hard negatives. arXiv preprint arXiv:1707.05612 (2017)
7. He, N., Liu, J., Zhang, B., Bui, N., Maatouk, A., Yang, M., King, I., Weber, M., Ying, R.: Position: Beyond euclidean–foundation models should embrace non-euclidean geometries. arXiv preprint arXiv:2504.08896 (2025)
8. Huang, J., Chen, Y., Xiong, S., Lu, X.: Visual contextual semantic reasoning for cross-modal drone image-text retrieval. *IEEE Transactions on Geoscience and Remote Sensing* (2024)
9. Huang, Z., Van Gool, L.: A Riemannian network for SPD matrix learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 31 (2017)
10. Khattak, M.U., Rasheed, H., Maaz, M., Khan, S., Khan, F.S.: Maple: Multi-modal prompt learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 19113–19122 (2023)
11. Kobler, R.J., Hirayama, J.i., Kawanabe, M.: Controlling the fréchet variance improves batch normalization on the symmetric positive definite manifold. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 3863–3867. IEEE (2022)
12. Liu, J., Jiang, S., Ge, W., Huang, W., Guo, B., Li, Q.: UAVPairs: A benchmark for match pair retrieval of large-scale UAV images. *ISPRS Journal of Photogrammetry and Remote Sensing* **230**, 227–240 (2025)
13. Liu, S., Chen, J., Pan, L., Ngo, C.W., Chua, T.S., Jiang, Y.G.: Hyperbolic visual embedding learning for zero-shot recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9273–9281 (2020)
14. Lu, H., Huo, Y., Yang, G., Lu, Z., Zhan, W., Tomizuka, M., Ding, M.: Uniadapter: Unified parameter-efficient transfer learning for cross-modal modeling. In: The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7–11, 2024. OpenReview.net (2024), <https://openreview.net/forum?id=f5H8WGLQm5>
15. Lu, X., Wang, B., Zheng, X., Li, X.: Exploring models and data for remote sensing image caption generation. *IEEE Transactions on Geoscience and Remote Sensing* **56**(4), 2183–2195 (2017)
16. Ma, Q., Pan, J., Bai, C.: Direction-oriented visual–semantic embedding model for remote sensing image–text retrieval. *IEEE Transactions on Geoscience and Remote Sensing* **62**, 1–14 (2024)
17. Monti, F., Boscaini, D., Masci, J., Rodola, E., Svoboda, J., Bronstein, M.M.: Geometric deep learning on graphs and manifolds using mixture model cnns. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5115–5124 (2017)
18. Ouyang, P., Ma, Q., Bai, C.: Sparse information perception network for remote sensing cross-modal retrieval. *IEEE Transactions on Geoscience and Remote Sensing* (2025)
19. Pal, A., Van Spengler, M., di Melendugno, G.M.D., Flaborea, A., Galasso, F., Mettes, P.: Compositional entailment learning for hyperbolic vision-language models. arXiv preprint arXiv:2410.06912 (2024)
20. Pan, J., Ma, Q., Bai, C.: A prior instruction representation framework for remote sensing image-text retrieval. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 611–620 (2023)

21. Pennec, X., Fillard, P., Ayache, N.: A Riemannian framework for tensor computing. *International Journal of Computer Vision* **66**(1), 41–66 (2006)
22. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: *Proceedings of the International Conference on Machine Learning*. pp. 8748–8763. PMLR (2021)
23. Ruan, H., Lin, J., Lai, Y., Luo, Z., Li, S.: HCCM: Hierarchical cross-granularity contrastive and matching learning for natural language-guided drones (2025)
24. Song, Y., Soleymani, M.: Polysemous visual-semantic embedding for cross-modal retrieval. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 1979–1988 (2019)
25. Wang, R., Jin, S., Chen, Z., Luo, X., Wu, X.J.: Learning to normalize on the SPD manifold under Bures–Wasserstein geometry. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 8289–8298 (2025)
26. Wang, R., Wu, X.J., Xu, T., Hu, C., Kittler, J.: Deep metric learning on the SPD manifold for image set classification. *IEEE Transactions on Circuits and Systems for Video Technology* **34**(2), 663–680 (2022)
27. Wu, Y., Li, L., Jiao, L., Liu, F., Liu, X., Yang, S.: TRTR-CMR: Cross-modal reasoning dual transformer for remote sensing image captioning. *IEEE Transactions on Geoscience and Remote Sensing* (2024)
28. Yang, R., Wang, S., Han, Y., Li, Y., Zhao, D., Quan, D., Guo, Y., Jiao, L., Yang, Z.: Transcending fusion: A multi-scale alignment method for remote sensing image-text retrieval. *IEEE Transactions on Geoscience and Remote Sensing* (2024)
29. Yang, X., Li, C., Wang, Z., Xie, H., Mao, J., Yin, G.: Remote sensing cross-modal text-image retrieval based on attention correction and filtering. *Remote Sensing* **17**(3), 503 (2025)
30. Yuan, Z., Zhang, W., Fu, K., Li, X., Deng, C., Wang, H., Sun, X.: Exploring a fine-grained multiscale method for cross-modal remote sensing image retrieval (2022)
31. Yuan, Z., Zhang, W., Tian, C., Rong, X., Zhang, Z., Wang, H., Fu, K., Sun, X.: Remote sensing cross-modal text-image retrieval based on global and local information. *IEEE Transactions on Geoscience and Remote Sensing* **60**, 1–16 (2022)