GUIDE: GEOMETRY-GROUNDED IDENTITY STRUCTURE LEARNING FOR MULTI-MODAL OBJECT RE-IDENTIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-modal object re-identification (ReID) aims to retrieve the same target across cameras by exploiting complementary cues from heterogeneous modalities. Recent graph-based approaches have further advanced multi-modal ReID by explicitly modeling relational dependencies among local regions and semantic representations. However, graph-based methods typically derive relational topology from feature or semantic similarity, causing corresponding regions across modalities to aggregate different contextual information during relational propagation, thereby weakening the consistency of identity-related structural cues. To address this issue, we propose GUIDE, a Geometry-Grounded Identity Structure Learning framework for multi-modal object ReID. GUIDE decouples relational topology from modality-dependent appearance by grounding connectivity in patch geometry while preserving modality-specific representations. Specifically, the Grid Graph Encoder (GGE) restores patch tokens to their native two-dimensional layouts and propagates local identity cues over geometry-defined neighborhoods using a graph encoder shared across modalities. The Axial Strip Readout (ASR) preserves spatial composition by aggregating graph-refined tokens into ordered strips with learnable queries. Together, GGE and ASR produce structured identity representations that complement global semantic features for retrieval. Extensive experiments on RGBNT201, RGBNT100, and MSVR310 demonstrate the effectiveness of GUIDE across person and vehicle ReID benchmarks.

1 INTRODUCTION

Object re-identification (ReID) aims to retrieve the same target across non-overlapping camera views. Although RGB-based ReID has achieved remarkable progress with advanced representation learning architectures (He et al., 2021; Li et al., 2023b; Zhang et al., 2023; Li et al., 2023a), its reliability can degrade under challenging imaging conditions such as illumination variation, shadows, low visibility, and sensor noise. Multi-modal object ReID alleviates these limitations by jointly exploiting heterogeneous sensing modalities, including RGB, near-infrared (NIR), and thermal infrared (TIR). These modalities respond differently to illumination, material properties, and thermal radiation, thereby providing complementary identity cues for robust retrieval under complex visual conditions (Li et al., 2020; Wang et al., 2022; 2024b).

Recent multi-modal ReID methods have mainly focused on improving *what information is represented*. Representative directions include cross-modal interaction, token-level representation learning, semantic guidance, modality-aware fusion, and feature decoupling (Wang et al., 2024a; Zhang et al., 2024; Feng et al., 2025; Wang et al., 2025b). These approaches substantially improve the utilization of heterogeneous visual cues. More recently, graph-based methods have introduced relational modeling into multi-modal ReID. STMI (Xu et al., 2026) constructs a unified cross-modal hypergraph based on semantic similarity to capture high-order interactions, while MGRNet (Wan et al., 2026) builds modality-specific graphs from local feature relationships for fine-grained reasoning. Although effective, both approaches derive relational structures from learned visual or semantic representations.

Figure 1 illustrates the key difference in how relational topology is constructed. As shown in Fig. 1(a), cross-modal similarity graphs connect regions according to similarities between heterogeneous

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

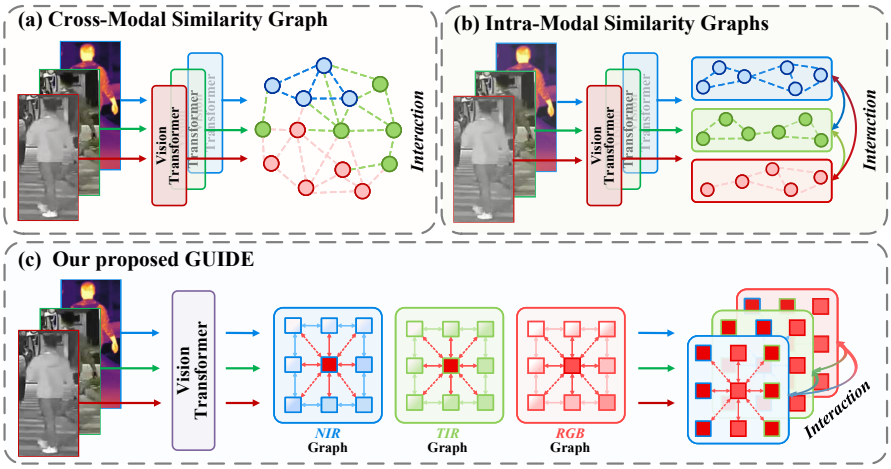


Figure 1: **Motivation and core principle of GUIDE.** (a) Cross-modal similarity graphs derive relational topology from similarities across heterogeneous representations, coupling node semantics with connectivity. (b) Intra-modal similarity graphs construct modality-specific neighborhoods independently, but their topology remains dependent on modality-specific representations and can vary across modalities. (c) Our GUIDE decouples relational topology from modality-dependent appearance by grounding connectivity in a common geometric rule while preserving modality-specific node representations, providing a shared structural basis for multi-modal identity learning.

representations, enabling relational interaction across modalities. In contrast, Fig. 1(b) constructs separate similarity graphs within each modality, where neighborhoods are determined independently from modality-specific features. Although the two paradigms differ in where relational modeling is performed, both rely on learned representations to determine graph connectivity. Under heterogeneous sensing, structurally corresponding regions in RGB, NIR, and TIR may produce substantially different feature responses. These differences can change their message-passing neighborhoods, causing corresponding regions to aggregate different contextual information during relational propagation. In other words, the same modality-dependent representations determine both node semantics and interaction paths, resulting in representation-topology coupling and weakening the consistency of identity-related structural cues across modalities.

To address this issue, we propose GUIDE, a Geometry-Grounded Identity Structure Learning framework for multi-modal object ReID. As shown in Fig. 1(c), GUIDE determines local connectivity from the native spatial geometry of patch tokens rather than feature similarity, while preserving modality-specific visual information in the node representations. This design provides a common relational support across modalities without forcing heterogeneous appearance features to be identical. Based on this principle, GUIDE models identity structure from two complementary aspects, local relational connectivity and spatial composition. Specifically, the Grid Graph Encoder (GGE) propagates identity cues over geometry-defined patch neighborhoods constructed with the same spatial rule across modalities, while the Axial Strip Readout (ASR) aggregates graph-refined tokens into ordered strips along the long axis of the native patch layout to preserve spatial composition. Together, GGE and ASR transform local patch representations into structured identity cues that complement the global semantic representation.

The main contributions are summarized as follows:

- We propose GUIDE, a Geometry-Grounded Identity Structure Learning framework for multi-modal object ReID. It addresses representation-topology coupling by separating modality-specific representations from geometry-grounded connectivity, providing a common geometric basis for relational modeling across modalities.
- We introduce a Grid Graph Encoder (GGE) and an Axial Strip Readout (ASR) to model identity structure from complementary relational and spatial aspects. GGE propagates local identity cues over geometry-defined neighborhoods constructed with the same spatial rule across modalities, while ASR preserves the spatial organization of graph-refined features during structural aggregation.

- Extensive experiments on RGBNT201, RGBNT100, and MSVR310 demonstrate GUIDE's effectiveness across person and vehicle ReID. Further analyses validate geometry-grounded topology and structure-preserving aggregation for multi-modal identity representation.

2 Related Work

Multi-modal object ReID has progressively evolved from cross-modal feature fusion toward fine-grained representation learning and relational modeling. Early studies exploit complementary information from heterogeneous modalities through feature interaction and fusion. With the emergence of Vision Transformers, subsequent approaches increasingly operate at the token level to improve cross-modal exchange and suppress irrelevant responses. TOP-ReID (Wang et al., 2024a) introduces cyclic token permutation to enhance tri-modal interaction, while EDITOR (Zhang et al., 2024) selects informative tokens to reduce background interference. More recent approaches further improve representation quality through modality decoupling, adaptive feature weighting, semantic guidance, and local feature modeling. DeMo (Wang et al., 2025a) adaptively balances decoupled modality representations with a mixture-of-experts, IDEA (Wang et al., 2025b) integrates semantic guidance with deformable aggregation for discriminative local feature learning, and MDReID (Feng et al., 2026) disentangles modality-shared and modality-specific representations to support flexible multi-modal retrieval. STMI (Xu et al., 2026) further incorporates segmentation-guided feature modulation and semantic token reallocation to enhance semantic representations before relational interaction. Beyond representation refinement, graph-based approaches explicitly model dependencies among local or semantic features. GPFNet (He et al., 2023) introduces graph-based progressive fusion for multi-modal interaction, while MGRNet (Wan et al., 2026) constructs modality-specific graphs from local feature relationships for fine-grained reasoning. STMI (Xu et al., 2026) constructs a unified cross-modal hypergraph whose hyperedges are determined by semantic similarity among semantic tokens. Although differing in node representation and propagation mechanism, these approaches demonstrate the effectiveness of relational modeling beyond independent feature encoding.

A common characteristic of these graph-based formulations is that connectivity is largely inferred from learned feature or semantic representations. Thus, the representations describing local content also influence the support over which information is propagated. Since graph topology determines which regions exchange information, deriving connectivity from modality-dependent similarity can make relational neighborhoods sensitive to modality-specific feature responses. GUIDE instead treats topology construction as a distinct aspect of relational modeling. Rather than using modality-sensitive similarity to define connectivity, GUIDE employs the native spatial layout of patch tokens as a common structural reference. The same geometric rule is applied across modalities while node representations remain modality-specific, providing shared geometric support for message passing without requiring appearance similarity to determine connectivity. Moreover, unlike conventional local or part-based methods that directly aggregate backbone features or selected tokens, GUIDE performs structural readout after geometry-grounded relational propagation, preserving the ordered spatial composition of graph-refined local cues while complementing the global semantic representation.

3 Methodology

We propose GUIDE, a Geometry-Grounded Identity Structure Learning framework for multi-modal object ReID that decouples relational topology from modality-specific feature responses. As illustrated in Fig. 2, GUIDE employs a CLIP-pretrained vanilla ViT-base shared across modalities and introduces the Grid Graph Encoder (GGE) and Axial Strip Readout (ASR) to model identity structure from relational and spatial perspectives. Given RGB, NIR, and TIR inputs, the encoder extracts global and patch tokens for each modality. GGE restores patch tokens to their native two-dimensional layouts and constructs modality-specific intra-modal graphs using a shared deterministic geometric rule, enabling local identity cues to propagate over geometry-defined neighborhoods while preserving modality-specific representations. Based on the graph-refined features, ASR organizes patch tokens into ordered strips along the long axis of the layout and summarizes each strip with a learnable query to preserve coarse spatial organization during aggregation. The resulting structural descriptor complements the global semantic representation. Finally, GUIDE combines global and structural representations for identity learning and retrieval, preserving heterogeneous appearance information while maintaining geometry-grounded support for relational propagation across modalities.

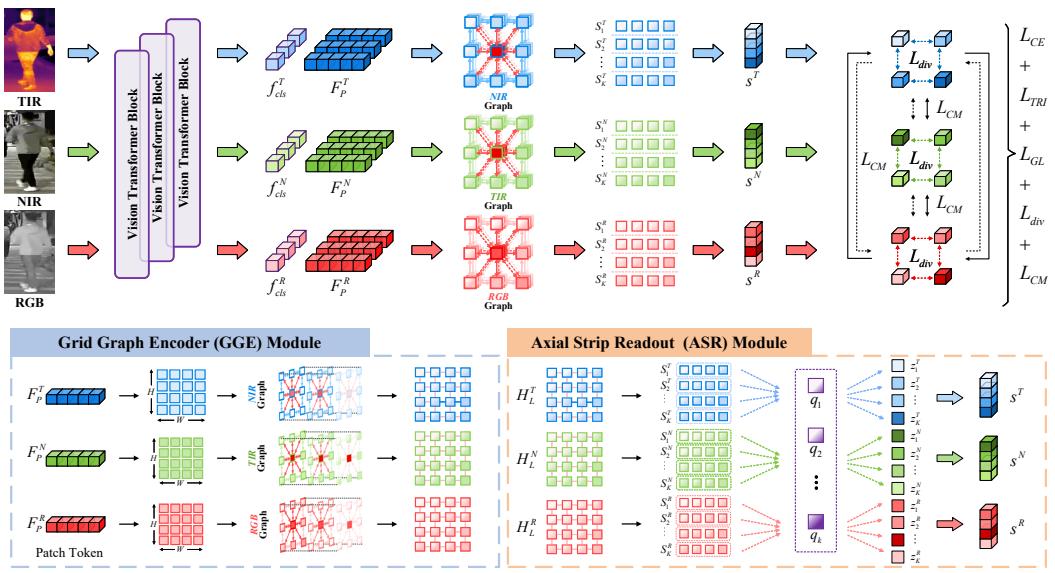


Figure 2: **Illustration of the proposed GUIDE framework.** The upper part presents the overall pipeline, where a shared Vision Transformer extracts global and patch-level representations from heterogeneous modalities. GGE builds geometry-grounded relational connectivity with shared topology, while ASR preserves spatial composition by aggregating graph-refined tokens into ordered strips with shared queries. The lower part details GGE and ASR, including geometry-defined propagation and query-guided structural readout. The final objective jointly optimizes identity discrimination, global–structural alignment, cross-modal consistency, and intra-modal strip diversity.

3.1 Grid Graph Encoder

Graph-based multi-modal ReID methods derive relational topology from feature or semantic similarity, implicitly allowing modality-dependent representations to determine both node semantics and connectivity. However, such similarity-driven connectivity is coupled with modality-dependent feature responses, causing the neighborhoods used for message passing to vary across heterogeneous modalities. To decouple relational connectivity from modality-specific appearance, we introduce the Grid Graph Encoder (GGE), which constructs geometry-grounded neighborhoods from the native spatial layout of patch tokens. Specifically, GGE preserves modality-specific patch representations while enforcing a shared spatial connectivity rule across RGB, NIR, and TIR modalities, thereby providing a consistent structural basis for relational modeling across heterogeneous observations.

Generally, GUIDE employs a CLIP-pretrained vanilla ViT-Base as the shared visual backbone across modalities. Given input images from different modalities ($M \in \{R, N, T\}$, corresponding to RGB, NIR, and TIR), each image is divided into non-overlapping patches following the standard ViT procedure and linearly projected into token embeddings as:

$$F^M = \left\{ \left[f^M_{\text{cls}} \right], f^M_1, f^M_2, \ldots, f^M_N \right\}, \qquad M \in \{R, N, T\}, \tag{1}$$

where $f^M_{\text{cls}} \in \mathbb{R}^D$ denotes the global class token, $f^M_i \in \mathbb{R}^D$ represents the i-th patch token, and D is the embedding dimension.

GUIDE retains both the global class token and patch tokens to capture complementary semantic and structural cues. Specifically, f^M_{cls} is preserved as the global semantic representation, while the patch tokens are formulated as:

$$F^M_p = \left\{ f^M_1, f^M_2, \ldots, f^M_N \right\} \in \mathbb{R}^{N \times D}, \qquad M \in \{R, N, T\}. \tag{2}$$

Although ViT processes patch tokens as a one-dimensional sequence, each token originates from a fixed spatial position in the original image plane. Therefore, GGE restores the native two-dimensional patch organization without modifying the token features. For an image divided into $H_p \times W_p$ patches,

the spatial coordinates of patch tokens can be explicitly organized as:

$$P = \begin{bmatrix} p_1 & p_2 & \cdots & p_{W_p} \\ p_{W_p+1} & p_{W_p+2} & \cdots & p_{2W_p} \\ \vdots & \vdots & \ddots & \vdots \\ p_{(H_p-1)W_p+1} & p_{(H_p-1)W_p+2} & \cdots & p_{H_pW_p} \end{bmatrix}, \tag{3}$$

where each position is defined as $p_i = (r_i, c_i)$, with r_i and c_i denoting the row and column index of the i-th patch on the native patch grid.

Accordingly, the patch tokens are reorganized into their native two-dimensional layout:

$$X^M = \mathrm{Reshape}(F_p^M) \in \mathbb{R}^{H_p \times W_p \times D}. \tag{4}$$

This operation only restores the spatial organization of patch tokens without changing their feature representations. The recovered patch layout therefore provides a common geometric reference for constructing local relational neighborhoods. Based on the restored patch layout, GGE constructs an independent intra-modal graph for each modality:

$$\mathcal{G}^M = (\mathcal{V}^M, \mathcal{E}^{geo}), \qquad M \in \{R, N, T\}, \tag{5}$$

where each patch token corresponds to a graph node:

$$\mathcal{V}^M = \{v_1^M, v_2^M, \ldots, v_N^M\}. \tag{6}$$

The geometric adjacency is determined according to the native patch coordinates. Specifically, we adopt an eight-connected neighborhood on the patch grid:

$$A_{ij}^{geo} = \mathbb{I}(0 < \|p_i - p_j\|_\infty \leq 1), \tag{7}$$

where $p_i = (r_i, c_i)$ denotes the spatial coordinate of the i-th patch token.

Unlike similarity-based graph construction, the connectivity is independent of modality-specific feature responses. Therefore, all modalities share the same geometric topology:

$$A^R = A^N = A^T = A^{geo}. \tag{8}$$

The initial node features are defined as:

$$H_0^M = F_p^M. \tag{9}$$

Although modalities share the same geometric topology, node features remain modality-specific:

$$H_0^R \neq H_0^N \neq H_0^T. \tag{10}$$

Before propagation, self-loops are added to the geometric adjacency matrix, followed by row normalization:

$$\tilde{A}^{geo} = A^{geo} + I, \tag{11}$$

$$\tilde{D}_{ii} = \sum_j \tilde{A}_{ij}^{geo}, \tag{12}$$

$$\hat{A}^{geo} = \tilde{D}^{-1}\tilde{A}^{geo}. \tag{13}$$

At the l-th graph layer, we first normalize the node features as $\bar{H}_l^M = \mathrm{LN}(H_l^M)$. The graph refinement is then formulated as:

$$H_{l+1}^M = H_l^M + \sigma\left(\hat{A}^{geo}\bar{H}_l^M W_l^m + \bar{H}_l^M W_l^s\right) W_l^o, \qquad l = 0, \ldots, L-1. \tag{14}$$

where W_l^m, W_l^s, and W_l^o denote the learnable message transformation, self-transformation, and output projection matrices.

Although the initial geometric edges only describe local spatial relationships, multiple refinement layers progressively expand the aggregation range through multi-hop message passing, allowing each node to capture broader identity-related contextual information. Since both the geometric connectivity and graph transformation parameters are shared across modalities, GGE establishes a modality-consistent structural reasoning process while maintaining modality-specific visual semantics.

3.2 AXIAL STRIP READOUT

Although GGE establishes geometry-grounded relational connectivity and refines local patch representations through message propagation, directly aggregating all graph-refined tokens into a single global descriptor may weaken their spatial organization. To preserve coarse spatial composition of identity-related local cues, we introduce the Axial Strip Readout (ASR), performing ordered structural aggregation over geometry-refined patch representations. Unlike order-agnostic feature pooling, ASR explicitly maintains the spatial arrangement of local relational cues during structural readout.

Given the final geometry-refined patch features $H_L^M \in \mathbb{R}^{N\times D}$ produced by GGE, ASR first restores them into the same native patch layout used in GGE:

$$Y^M = \mathrm{Reshape}(H_L^M) \in \mathbb{R}^{H_p\times W_p\times D}, \qquad M \in \{R, N, T\}, \tag{15}$$

where H_p and W_p denote the height and width of the patch grid, respectively. This restoration only reorganizes the graph-refined tokens according to their original spatial positions while preserving the learned relational features.

To retain the coarse spatial composition of local identity cues, ASR organizes the graph-refined tokens into ordered strips along the long axis of the native patch layout. Specifically, the restored feature layout is evenly divided into K ordered strips along its longer spatial dimension:

$$S^M = [S_1^M, S_2^M, \ldots, S_K^M], \tag{16}$$

where each strip contains consecutive patch tokens along the long axis. The ordering of strips is preserved throughout aggregation, enabling ASR to maintain coarse spatial composition without requiring explicit semantic part annotations.

For each strip, ASR introduces a learnable query vector $q_k \in \mathbb{R}^D$ to adaptively aggregate informative local features. To ensure a consistent structural readout across modalities, the strip queries are shared among RGB, NIR, and TIR:

$$q_k^R = q_k^N = q_k^T = q_k. \tag{17}$$

Given the i-th token $h_{k,i}^M$ within the k-th strip, ASR computes its contribution weight based on the similarity between the learnable strip query and the token feature, where the summation is performed over all tokens within the k-th strip:

$$\alpha_{k,i}^M = \frac{\exp((q_k)^T h_{k,i}^M)}{\sum_j \exp((q_k)^T h_{k,j}^M)}, \qquad M \in \{R, N, T\}. \tag{18}$$

The structural representation of each strip is obtained through query-guided aggregation:

$$z_k^M = \sum_i \alpha_{k,i}^M h_{k,i}^M. \tag{19}$$

Finally, the ordered strip representations are concatenated and projected into the structural descriptor:

$$s^M = \phi_f\left(\mathrm{Concat}\left(z_1^M, \ldots, z_K^M\right)\right), \tag{20}$$

where $\phi_f(\cdot)$ denotes the readout projection with layer normalization and linear transformation.

Maintaining strip representation order, ASR preserves coarse axial organization of graph-refined local cues while adaptively emphasizing informative regions. The resulting structural descriptor complements the global semantic representation by providing ordered structural cues for identity learning and retrieval.

3.3 LEARNING OBJECTIVE

As shown in Fig. 2, we optimize the framework using multiple losses. The learned identity features are supervised by the label-smoothed cross-entropy loss and triplet loss:

$$\mathcal{L}_g = \mathcal{L}_{CE} + \mathcal{L}_{TRI}, \tag{21}$$

where $\mathcal{L}_{CE}$ and $\mathcal{L}_{TRI}$ denote CE and triplet losses, and $\mathcal{L}_g$ supervises both representations.

Besides identity discrimination, we introduce structural regularization to constrain representations learned by GGE and ASR. For sample b, we denote concatenated global and structural representations as $g_b = \mathrm{Concat}(f^R_{\mathrm{cls},b}, f^N_{\mathrm{cls},b}, f^T_{\mathrm{cls},b})$ and $s_b = \mathrm{Concat}(s^R_b, s^N_b, s^T_b)$, respectively. To encourage the structural descriptor to retain identity semantics compatible with global representation, we employ a global-structural consistency loss over a batch of B samples using cosine similarity $\cos(\cdot)$:

$$\mathcal{L}_{\mathrm{GL}} = \frac{1}{B}\sum_{b=1}^{B}\left[1-\cos\left(g_b, s_b\right)\right], \tag{22}$$

Although ASR preserves the ordered spatial composition of local identity cues, different strip representations may converge to redundant responses. We therefore introduce a strip diversity loss to encourage complementary structural information among different strips:

$$\mathcal{L}_{\mathrm{div}} = \frac{1}{3BK(K-1)}\sum_{M\in\{R,N,T\}}\sum_{b=1}^{B}\sum_{i\neq j}\max\left(0, \cos\left(z^M_{b,i}, z^M_{b,j}\right) - m\right), \tag{23}$$

where $z^M_{b,i}$ denotes the representation of the i-th strip from modality M for the b-th sample, and m denotes the diversity margin.

Since GGE and ASR employ shared structural modeling across modalities, we consider the modality pairs $\mathcal{P} = \{(R,N),(R,T),(N,T)\}$ and further encourage corresponding RGB, NIR, and TIR observations to maintain consistent structural identity information. The cross-modal structural consistency loss is defined as:

$$\mathcal{L}_{\mathrm{CM}} = \frac{1}{3B}\sum_{b=1}^{B}\sum_{(M_i,M_j)\in\mathcal{P}}\left[1-\cos\left(s^{M_i}_b, s^{M_j}_b\right)\right]. \tag{24}$$

This constraint is imposed only on the final structural descriptors, rather than on the modality-specific node representations, thereby encouraging identity-level structural consistency without eliminating modality-specific appearance characteristics. Finally, the overall objective of GUIDE is formulated as

$$\mathcal{L} = \mathcal{L}_g + \lambda_{\mathrm{GL}}\mathcal{L}_{\mathrm{GL}} + \lambda_{\mathrm{div}}\mathcal{L}_{\mathrm{div}} + \lambda_{\mathrm{CM}}\mathcal{L}_{\mathrm{CM}}, \tag{25}$$

where λ_{GL}, λ_{div}, and λ_{CM} control the contributions of the corresponding objectives.

During inference, the global and structural representations are independently normalized and combined through a weighted addition. The fused representation is further normalized and used as the final representation for retrieval.

4 Experiments

4.1 Datasets and Evaluation Protocols

We evaluate the proposed **GUIDE** framework on three multi-modal object ReID benchmarks, including the person ReID dataset RGBNT201 (Zheng et al., 2021) and the vehicle ReID datasets RGBNT100 (Li et al., 2020) and MSVR310 (Zheng et al., 2023b). Performance is evaluated using mean Average Precision (mAP) and Cumulative Matching Characteristics (CMC) at ranks 1, 5, and 10 (R-1, R-5, and R-10). We further assess robustness under missing-modality conditions on RGBNT201, where one or multiple modalities are unavailable during inference. RGBNT201 contains 201 identities with 4,787 images per modality, while RGBNT100 and MSVR310 provide 17,250 and 2,087 aligned multi-modal vehicle samples from 100 and 310 vehicles, respectively.

4.2 Implementation Details

All experiments are implemented in PyTorch and conducted on a single NVIDIA RTX 3090 GPU with CUDA 12.5 and Python 3.8. We adopt the CLIP-Base visual encoder (Radford et al., 2021) as the backbone. Input images are resized to 256×128 for RGBNT201 and 128×256 for RGBNT100 and MSVR310. Random flipping, cropping, and random erasing (Zhong et al., 2020) are used for augmentation. The model is optimized with Adam, batch size 64. The initial learning rate is 3.5×10^{-4}, while the visual encoder is fine-tuned with 5×10^{-6}. Training runs for 50 epochs.

Table 1: **Performance comparison on RGBNT201, RGBNT100, and MSVR310.** The best and second results are in bold and underlined, respectively.

Methods	RGBNT201				Methods	RGBNT100		MSVR310	
	*m*AP	R-1	R-5	R-10		*m*AP	R-1	*m*AP	R-1
Single									
MUDeep (Qian et al., 2017)	23.8	19.7	33.1	44.3	DMML (Chen et al., 2019)	58.5	82.0	19.1	31.1
HACNN (Li et al., 2018)	21.3	19.0	34.1	42.8	BoT (Luo et al., 2019)	78.0	95.1	23.5	38.4
MLFN (Chang et al., 2018)	26.1	24.2	35.9	44.1	Circle Loss (Sun et al., 2020)	59.4	81.7	22.7	34.2
PCB (Sun et al., 2018)	32.8	28.1	37.4	46.9	HRCN (Zhao et al., 2021)	67.1	91.8	23.4	44.2
OSNet (Zhou et al., 2019)	25.4	22.3	35.1	44.7	TransReID (He et al., 2021)	75.6	92.9	18.4	29.6
CAL (Rao et al., 2021)	27.6	24.3	36.5	45.7	AGW (Ye et al., 2022)	73.1	92.7	28.9	46.9
Multi									
PFNet (Zheng et al., 2021)	38.5	38.9	52.0	58.4	GraFT (Yin et al., 2023)	76.6	94.3	-	-
IEEE (Wang et al., 2022)	47.5	44.4	57.1	63.6	GPFNet (He et al., 2023)	75.0	94.5	-	-
DENet (Zheng et al., 2023a)	42.4	42.2	55.3	64.5	PHT (Pan et al., 2023)	79.9	92.7	-	-
UniCat (Crawford et al., 2023)	57.0	55.7	-	-	UniCat (Crawford et al., 2023)	79.4	96.2	-	-
HTT (Wang et al., 2024b)	71.1	73.4	83.1	87.3	CCNet (Zheng et al., 2023b)	77.2	96.3	36.4	55.2
EDITOR (Zhang et al., 2024)	66.5	68.3	81.1	88.2	EDITOR (Zhang et al., 2024)	82.1	96.4	39.0	49.3
RSCNet (Yu et al., 2024)	68.2	72.5	-	-	RSCNet (Yu et al., 2024)	82.3	96.6	39.5	49.6
TOP-ReID (Wang et al., 2024a)	72.3	76.6	84.7	89.4	TOP-ReID (Wang et al., 2024a)	81.2	96.4	35.9	44.6
LRMM (Wu et al., 2025)	52.3	53.4	64.6	73.2	LRMM (Wu et al., 2025)	78.6	96.7	36.7	49.7
WTSF-ReID (Yu et al., 2025)	67.9	72.2	83.4	89.7	WTSF-ReID (Yu et al., 2025)	82.2	96.5	39.2	49.1
DESANet (Dong et al., 2025)	74.6	77.6	87.1	91.3	DESANet (Dong et al., 2025)	82.1	97.4	39.2	47.8
DeMo (Wang et al., 2025a)	79.0	82.3	88.8	92.0	DeMo (Wang et al., 2025a)	86.2	**97.6**	49.2	59.8
IDEA (Wang et al., 2025b)	80.2	82.1	90.0	93.3	IDEA (Wang et al., 2025b)	87.2	96.5	47.0	62.4
MDReID (Feng et al., 2026)	82.1	**85.2**	90.3	92.6	MDReID (Feng et al., 2026)	85.3	95.6	51.0	68.9
STMI (Xu et al., 2026)	81.2	83.4	90.2	91.6	MGRNet (Wan et al., 2026)	87.8	97.3	53.2	67.2
GUIDE (Ours)	**83.0**	85.0	**90.8**	**94.1**	**GUIDE (Ours)**	**88.1**	97.2	**53.7**	**69.4**

Table 2: **Performance of missing-modality settings on RGBNT201.** "M (X)" means missing the X image modality. The best and second results are in bold and underlined, respectively.

	Methods	M (RGB)		M (NIR)		M (TIR)		M (RGB+NIR)		M (RGB+TIR)		M (NIR+TIR)		Average	
		*m*AP	R-1	*m*AP	R-1	*m*AP	R-1	*m*AP	R-1	*m*AP	R-1	*m*AP	R-1	*m*AP	R-1
Single	MUDeep (Qian et al., 2017)	19.2	16.4	20.0	17.2	18.4	14.2	13.7	11.8	11.5	6.5	12.7	8.5	15.9	12.9
	HACNN (Li et al., 2018)	12.5	11.1	20.5	19.4	16.7	13.3	9.2	6.2	6.3	2.2	14.8	12.0	13.3	10.7
	MLFN (Chang et al., 2018)	20.2	18.9	21.1	19.7	17.6	11.1	13.2	12.1	8.3	3.5	13.1	9.1	15.6	12.4
	CAL (Rao et al., 2021)	21.4	22.1	24.2	23.6	18.0	12.4	18.6	20.1	10.0	5.9	17.2	13.2	18.2	16.2
	PCB (Sun et al., 2018)	23.6	24.2	24.4	25.1	19.9	14.7	20.6	23.6	11.0	6.8	18.6	14.4	19.7	18.1
	OSNet (Zhou et al., 2019)	19.8	17.3	21.0	19.0	18.7	14.6	12.3	10.9	9.4	5.4	13.0	10.2	15.7	12.9
Multi	PFNet (Zheng et al., 2021)	-	-	31.9	29.8	25.5	25.8	-	-	-	-	26.4	23.4	-	-
	DENet (Zheng et al., 2023a)	-	-	35.4	36.8	33.0	35.4	-	-	-	-	32.4	29.2	-	-
	TOP-ReID (Wang et al., 2024a)	54.4	57.5	64.3	67.6	51.9	54.5	35.3	35.4	26.2	26.0	34.1	31.7	44.4	45.4
	DESANet (Dong et al., 2025)	54.1	55.1	65.0	67.3	56.6	56.6	35.7	38.2	**28.2**	**28.8**	34.2	29.8	45.7	46.0
	GUIDE (Ours)	**62.9**	**62.7**	**72.9**	**77.0**	**60.5**	**59.0**	**42.6**	**42.7**	27.3	**28.8**	**42.9**	**43.2**	**51.5**	**52.2**

4.3 Comparison with State-of-the-Art Methods

In Table 1, we compare GUIDE with existing single-modal and multi-modal methods on RGBNT201, RGBNT100, and MSVR310. Multi-modal approaches consistently outperform single-modal methods, demonstrating the benefit of exploiting heterogeneous modality information for robust ReID. On RGBNT201, GUIDE achieves the best mAP and Rank-5/Rank-10 retrieval performance with 83.0% mAP, 85.0% Rank-1, 90.8% Rank-5, and 94.1% Rank-10 accuracy. Compared with TOP-ReID and IDEA, GUIDE improves mAP by 10.7% and 2.8%, respectively. It also surpasses STMI and MDReID by 1.8% and 0.9% in mAP, respectively. These improvements suggest that geometry-grounded topology provides a more stable structural basis for identity representation under heterogeneous sensing. On RGBNT100 and MSVR310, GUIDE achieves 88.1% and 53.7% mAP, respectively, demonstrating its generalization across both person and vehicle ReID benchmarks.

To evaluate robustness under incomplete observations, we conduct missing-modality experiments on RGBNT201, where one or multiple modalities are unavailable at inference. As shown in Table 2, GUIDE achieves the best average performance, with 51.5% mAP and 52.2% Rank-1 accuracy, improving over DESANet by 5.8% and 6.2%, respectively. GUIDE also demonstrates strong robustness across missing-modality settings, including severe cases with two modalities simultaneously unavailable. This robustness suggests that GUIDE does not rely solely on modality-specific appearance cues. Instead, GGE and ASR preserve modality-independent structural information through geometry-grounded connectivity and structure-preserving aggregation, enabling reliable retrieval under incomplete sensory inputs.

Table 3: **Ablation study and hyper-parameter settings of GUIDE.** The best performance is highlighted in bold.

(a) **Ablation study.** GGE and ASR denote the Grid Graph Encoder and Axial Strip Readout, while $\mathcal{L}_{\mathrm{GL}}$, $\mathcal{L}_{\mathrm{div}}$, and $\mathcal{L}_{\mathrm{CM}}$ denote the proposed structural regularization terms.

Index	GGE	ASR	$\mathcal{L}_{\mathrm{GL}}$	$\mathcal{L}_{\mathrm{div}}$	$\mathcal{L}_{\mathrm{CM}}$	mAP	R-1
1	×	×	×	×	×	76.0	77.8
2	✓	×	×	×	×	79.8	82.8
3	✓	✓	×	×	×	81.1	83.1
4	✓	✓	✓	×	×	81.5	83.3
5	✓	✓	✓	✓	×	82.0	83.4
6	✓	✓	✓	✓	✓	**83.0**	**85.0**

(b) **Comparison with different λ_{GL} for GUIDE.**

λ_{GL}	mAP	R-1
0.1	80.5	82.9
0.2	**81.5**	**83.3**
0.3	81.3	83.1

(c) **Comparison with different λ_{div} for GUIDE.**

λ_{div}	mAP	R-1
0.0	81.5	83.3
0.1	**82.0**	**83.4**
0.2	81.8	83.3

(d) **Comparison with different λ_{CM} for GUIDE.**

λ_{CM}	mAP	R-1
0.1	82.0	83.4
0.2	**83.0**	**85.0**
0.3	82.7	84.8

4.4 Ablation Study

We conduct ablation studies on RGBNT201 to evaluate the effectiveness of GGE, ASR, and three structural regularization terms, including the global-structural alignment loss $\mathcal{L}_{\mathrm{GL}}$, strip diversity loss $\mathcal{L}_{\mathrm{div}}$, and cross-modal consistency loss $\mathcal{L}_{\mathrm{CM}}$. For variants without ASR, graph-refined patch features are average-pooled to obtain the structural descriptor.

Effectiveness of GGE and ASR. Table 3(a) shows that the baseline without GGE and ASR achieves 76.0% mAP and 77.8% Rank-1 accuracy. Introducing GGE improves the performance to 79.8% mAP and 82.8% Rank-1 accuracy, demonstrating that geometry-grounded relational propagation provides more effective local structural reasoning with modality-consistent neighborhoods. Further incorporating ASR increases the performance to 81.1% mAP and 83.1% Rank-1 accuracy, indicating that preserving the spatial organization of graph-refined features introduces complementary structural cues beyond relational propagation.

Effectiveness of Structural Regularization. Starting from the model with GGE and ASR, adding $\mathcal{L}_{\mathrm{GL}}$ improves the mAP from 81.1% to 81.5%, validating aligning global semantic and structural representations. Introducing $\mathcal{L}_{\mathrm{div}}$ increases the mAP to 82.0%, demonstrating that encouraging diversity among strip representations helps reduce redundant structural responses. Finally, incorporating $\mathcal{L}_{\mathrm{CM}}$ yields the highest performance of 83.0% mAP and 85.0% Rank-1 accuracy by enhancing cross-modal structural consistency while preserving modality-specific appearance information.

Hyper-parameter Analysis. Tables 3(b)-(d) analyze the influence of structural regularization weights. The optimal settings are $\lambda_{\mathrm{GL}} = 0.2$ (81.5% mAP, 83.3% Rank-1 accuracy), $\lambda_{\mathrm{div}} = 0.1$ (82.0% mAP, 83.4% Rank-1 accuracy), and $\lambda_{\mathrm{CM}} = 0.2$ (83.0% mAP, 85.0% Rank-1 accuracy). These results indicate that moderate structural constraints provide effective regularization, whereas excessive weights may over-constrain consistency and suppress modality-specific discriminative information.

5 Conclusion

We propose **GUIDE**, a Geometry-Grounded Identity Structure Learning framework for multi-modal object ReID. GUIDE decouples relational topology from modality-dependent appearance by grounding connectivity in the native spatial geometry of patch tokens while preserving modality-specific representations. The Grid Graph Encoder (GGE) constructs shared geometry-defined neighborhoods across modalities and propagates local identity cues through relational reasoning. The Axial Strip Readout (ASR) preserves spatial composition by aggregating graph-refined tokens into ordered strips with shared queries, producing complementary structural representations. Extensive experiments on RGBNT201, RGBNT100, and MSVR310 demonstrate the effectiveness and generalization of GUIDE for robust multi-modal object ReID.

Ethics statement

This study does not involve human subjects, personally identifiable information, or sensitive user data. The experiments are conducted on publicly available datasets that have been widely used in prior research. We comply with the corresponding dataset licenses and usage agreements and do not identify any additional ethical concerns. The proposed methods are intended solely for academic research purposes and do not introduce foreseeable harmful applications.

Reproducibility statement

We made our best efforts to comprehensively document the implementation details. Model architectures and the proposed modules are described in Section 3, while training hyperparameters and implementation details are provided in Section 4.2. We include the dataset descriptions and evaluation protocols for all benchmarks in Section 4.1. For evaluation, we follow the official protocols of the corresponding benchmarks and report the full results across different datasets and missing-modality settings in the supplementary material.

References

Xiaobin Chang, Timothy M. Hospedales, and Tao Xiang. Multi-level factorisation net for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Guangyi Chen, Tianren Zhang, Jiwen Lu, and Jie Zhou. Deep meta metric learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

Jennifer Crawford, Haoli Yin, Luke McDermott, and Daniel Cummings. Unicat: Crafting a stronger fusion baseline for multimodal re-identification, 2023. URL https://arxiv.org/abs/2310.18812.

Wenjiao Dong, Xi Yang, De Cheng, Nannan Wang, and Xinbo Gao. Escaping modal interactions: An efficient desanet for multi-modal object re-identification. *IEEE Transactions on Image Processing*, 2025.

Yingying Feng, Jie Li, Chi Xie, Lei Tan, and Jiayi Ji. Multi-modal object re-identification via sparse mixture-of-experts. In *Forty-second International Conference on Machine Learning*, 2025.

Yingying Feng, Jie Li, Jie Hu, Yukang Zhang, Lei Tan, and Jiayi Ji. Mdreid: Modality-decoupled learning for any-to-any multi-modal object re-identification. *Advances in Neural Information Processing Systems*, 38:42584–42605, 2026.

Qiaolin He, Zefeng Lu, Zihan Wang, and Haifeng Hu. Graph-based progressive fusion network for multi-modality vehicle re-identification. *IEEE Transactions on Intelligent Transportation Systems*, 24(11):12431–12447, 2023. doi: 10.1109/TITS.2023.3285758.

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

Hongchao Li, Chenglong Li, Xianpeng Zhu, Aihua Zheng, and Bin Luo. Multi-spectral vehicle re-identification: A challenge. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):11345–11353, Apr. 2020. doi: 10.1609/aaai.v34i07.6796. URL https://ojs.aaai.org/index.php/AAAI/article/view/6796.

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.

Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

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.

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

Wenjie Pan, Linhan Huang, Jianbao Liang, Lan Hong, and Jianqing Zhu. Progressively hybrid transformer for multi-modal vehicle re-identification. *Sensors*, 23(9), 2023. ISSN 1424-8220. doi: 10.3390/s23094206. URL https://www.mdpi.com/1424-8220/23/9/4206.

Xuelin Qian, Yanwei Fu, Yu-Gang Jiang, Tao Xiang, and Xiangyang Xue. Multi-scale deep learning architectures for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Yongming Rao, Guangyi Chen, Jiwen Lu, and Jie Zhou. Counterfactual attention learning for fine-grained visual categorization and re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1025–1034, October 2021.

Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. Circle loss: A unified perspective of pair similarity optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Xixi Wan, Aihua Zheng, Zi Wang, Bo Jiang, Jin Tang, and Jixin Ma. Reliable multi-modal object re-identification via modality-aware graph reasoning. *IEEE Transactions on Information Forensics and Security*, 2026.

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, 2024a.

Yuhao Wang, Yang Liu, Aihua Zheng, and Pingping Zhang. Decoupled feature-based mixture of experts for multi-modal object re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 8141–8149, 2025a.

Yuhao Wang, Yongfeng Lv, Pingping Zhang, and Huchuan Lu. Idea: Inverted text with cooperative deformable aggregation for multi-modal object re-identification. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 29701–29710, 2025b.

Zi Wang, Chenglong Li, Aihua Zheng, Ran He, and Jin Tang. Interact, embed, and enlarge: Boosting modality-specific representations for multi-modal person re-identification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(3):2633–2641, Jun. 2022. doi: 10.1609/aaai.v36i3.20165. URL https://ojs.aaai.org/index.php/AAAI/article/view/20165.

Zi Wang, Huaibo Huang, Aihua Zheng, and Ran He. Heterogeneous test-time training for multi-modal person re-identification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38 (6):5850–5858, Mar. 2024b. doi: 10.1609/aaai.v38i6.28398. URL https://ojs.aaai.org/index.php/AAAI/article/view/28398.

Di Wu, Zhihui Liu, Zihan Chen, Shenglong Gan, Kaiwen Tan, Qin Wan, and Yaonan Wang. Lrmm: Low rank multi-scale multi-modal fusion for person re-identification based on rgb-ni-ti. *Expert Systems with Applications*, 263:125716, 2025.

Xingguo Xu, Zhanyu Liu, Weixiang Zhou, Yuansheng Gao, Junjie Cao, Yuhao Wang, Jixiang Luo, and Dell Zhang. Stmi: Segmentation-guided token modulation with cross-modal hypergraph interaction for multi-modal object re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 11433–11441, 2026.

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

Haoli Yin, Jiayao Li, Eva Schiller, Luke McDermott, and Daniel Cummings. Graft: Gradual fusion transformer for multimodal re-identification, 2023. URL https://arxiv.org/abs/2310.16856.

Zhi Yu, Zhiyong Huang, Mingyang Hou, Jiaming Pei, Yan Yan, Yushi Liu, and Daming Sun. Representation selective coupling via token sparsification for multi-spectral object re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. doi: 10.1109/TCSVT.2024.3509817.

Zhi Yu, Zhiyong Huang, Mingyang Hou, Yan Yan, and Yushi Liu. Wtsf-reid: Depth-driven window-oriented token selection and fusion for multi-modality vehicle re-identification with knowledge consistency constraint. *Expert Systems with Applications*, 274:126921, 2025.

Guiwei Zhang, Yongfei Zhang, Tianyu Zhang, Bo Li, and Shiliang Pu. Pha: Patch-wise high-frequency augmentation for transformer-based person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14133–14142, 2023.

Pingping Zhang, Yuhao Wang, Yang Liu, Zhengzheng Tu, and Huchuan Lu. Magic tokens: Select diverse tokens for multi-modal object re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 17117–17126, 2024.

Jiajian Zhao, Yifan Zhao, Jia Li, Ke Yan, and Yonghong Tian. Heterogeneous relational complement for vehicle re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 205–214, October 2021.

Aihua Zheng, Zi Wang, Zihan Chen, Chenglong Li, and Jin Tang. Robust multi-modality person re-identification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(4):3529–3537, May 2021. doi: 10.1609/aaai.v35i4.16467. URL https://ojs.aaai.org/index.php/AAAI/article/view/16467.

Aihua Zheng, Ziling He, Zi Wang, Chenglong Li, and Jin Tang. Dynamic enhancement network for partial multi-modality person re-identification. *arXiv preprint arXiv:2305.15762*, 2023a.

Aihua Zheng, Xianpeng Zhu, Zhiqi Ma, Chenglong Li, Jin Tang, and Jixin Ma. Multi-spectral vehicle re-identification with cross-directional consistency network and a high-quality benchmark, 2023b. URL https://arxiv.org/abs/2208.00632.

Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):13001–13008, Apr. 2020. doi: 10.1609/aaai.v34i07.7000. URL https://ojs.aaai.org/index.php/AAAI/article/view/7000.

Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Omni-scale feature learning for person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

A APPENDIX

AI USE STATEMENT

To improve the quality of the manuscript, Large Language Models (LLMs) were used to assist with writing and editing, including language polishing, grammar checking, and improving readability. The use of LLMs was strictly limited to linguistic refinement and did not involve generating research ideas, designing the proposed methodology, conducting experiments, analyzing results, or producing scientific claims. The authors carefully reviewed all LLM-assisted content and remain fully responsible for the final manuscript.

A.1 ANALYSIS OF GEOMETRY-GROUNDED RELATIONAL MODELING

Table 4: **Comparison with convolutional spatial modeling on RGBNT201.**

Methods	RGBNT201				
	mAP	R-1	R-5	R-10	Average
3×3 DWConv	78.9	81.3	87.3	90.3	84.5
GGE (Ours)	**83.0**	**85.0**	**90.8**	**94.1**	**88.2**

To distinguish GGE from conventional convolutional spatial modeling, we replace GGE with a 3×3 depth-wise convolutional block while keeping the remaining architecture and training settings unchanged. Both variants operate on the restored two-dimensional patch layout and exploit the same local spatial range. As shown in Table 4, GGE improves mAP from 78.9% to 83.0%. This comparison suggests that the benefit of GGE cannot be attributed solely to recovering two-dimensional locality. Different from convolutional operators that jointly perform local aggregation and feature transformation within predefined kernels, GGE explicitly decouples relational topology from feature transformation, allowing the propagation support to remain modality-consistent while preserving modality-specific identity cues.

A.2 VISUALIZATION ANALYSIS

A.2.1 MULTI-MODAL FEATURE DISTRIBUTIONS

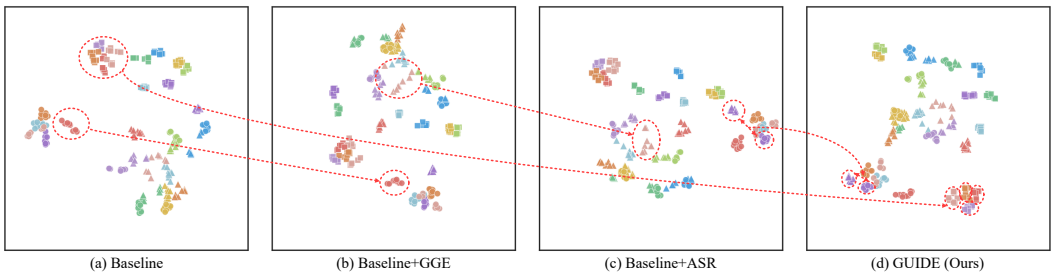


Figure 3: **t-SNE visualization of learned feature distributions under different model variants.** Different colors represent different identities. The four variants are Baseline, Baseline+GGE, Baseline+ASR, and GUIDE.

To qualitatively examine how the proposed structural components shape the learned representation space, we visualize the feature distributions of different model variants using t-SNE (Van der Maaten & Hinton, 2008). As shown in Fig. 3, different colors denote different identities. For the ASR-only variant, ASR is directly applied to the backbone patch tokens without graph refinement, allowing the contribution of spatially ordered structural readout to be examined independently. The baseline mainly relies on global semantic representations and exhibits relatively dispersed intra-class distributions and ambiguous inter-class boundaries in the t-SNE embedding. After introducing GGE, identity distributions appear more compact. This pattern is consistent with the hypothesis that geometry-grounded connectivity provides a more reliable basis for local structural representation learning by decoupling relational topology from modality-dependent feature similarity. In contrast, the ASR-only

variant also exhibits a more organized feature distribution, suggesting that preserving the spatial composition of patch-level cues provides structural information that is not fully captured by the global representation alone.

When GGE and ASR are combined in GUIDE, ASR operates on the graph-refined patch representations and preserves their ordered spatial composition during structural readout. GUIDE exhibits the clearest identity organization in the t-SNE embedding, with visually more compact intra-class groupings and clearer inter-class boundaries. Rather than serving as quantitative evidence of retrieval performance, these visualizations provide qualitative support for the central design of GUIDE. Specifically, GGE contributes geometry-grounded relational structure, while ASR preserves spatial composition, and their combination yields a structured identity representation that complements global semantic information.

A.2.2 Rank List Comparison



Figure 4: **Rank list comparison of different GUIDE variants on person ReID.** Baseline, Baseline+GGE, Baseline+ASR, and GUIDE are compared under the same query samples to illustrate their retrieval behavior across camera views. Green and red boxes indicate correct and incorrect identity matches, respectively.

To further analyze how different structural components influence retrieval behavior, we visualize the top-10 ranking lists of the baseline, baseline+GGE, baseline+ASR, and the full GUIDE model in Fig. 4. The baseline relies on global semantic representations, and several visually similar but identity-inconsistent samples appear in the top-ranked positions under cross-camera variations. This indicates that global appearance similarity alone may not fully capture fine-grained identity-related structural cues.

After introducing GGE, the ranking results become more reliable, with more correct identities retrieved at higher positions and fewer false matches among the top candidates. This improvement suggests that geometry-grounded relational propagation enhances local structural reasoning by exploiting spatially neighboring regions with consistent connectivity across modalities. The ASR-only variant also improves retrieval quality by preserving the spatial organization of patch-level identity cues. By maintaining the ordered composition of local features, ASR helps distinguish samples with similar global appearances but different spatial structures. When GGE and ASR are jointly integrated, GUIDE achieves the most reliable retrieval results, with correct identities consistently appearing at earlier ranks. These observations demonstrate that relational propagation and structure-preserving aggregation provide complementary benefits for constructing discriminative identity representations.