

UGG-ReID: Uncertainty-Guided Graph Model for Multi-Modal Object Re-Identification

Xixi Wan¹ Aihua Zheng^{1*} Bo Jiang^{2*} Beibei Wang² Chenglong Li¹ Jin Tang²

¹Information Materials and Intelligent Sensing Laboratory of Anhui Province,
School of Artificial Intelligence, Anhui University, Hefei, China

²Anhui Provincial Key Laboratory of Multimodal Cognitive Computation,
School of Computer Science and Technology, Anhui University, Hefei, China

¹ahzheng214@foxmail.com, ²jiangbo@ahu.edu.cn

Abstract

Multi-modal object Re-IDentification (ReID) has gained considerable attention with the goal of retrieving specific targets across cameras using heterogeneous visual data sources. At present, multi-modal object ReID faces two core challenges: (1) learning robust features under fine-grained local noise caused by occlusion, frame loss, and other disruptions; and (2) effectively integrating heterogeneous modalities to enhance multi-modal representation. To address the above challenges, we propose a robust approach named Uncertainty-Guided Graph model for multi-modal object ReID (UGG-ReID). UGG-ReID is designed to mitigate noise interference and facilitate effective multi-modal fusion by **estimating both local and sample-level aleatoric uncertainty and explicitly modeling their dependencies**. Specifically, we first propose the Gaussian patch-graph representation model that leverages uncertainty to quantify fine-grained local cues and capture their structural relationships. This process boosts the expressiveness of modal-specific information, ensuring that the generated embeddings are both more informative and robust. Subsequently, we design an uncertainty-guided mixture of experts strategy that dynamically routes samples to experts exhibiting low uncertainty. This strategy effectively suppresses noise-induced instability, leading to enhanced robustness. Meanwhile, we design an uncertainty-guided routing to strengthen the multi-modal interaction, improving the performance. UGG-ReID is comprehensively evaluated on five representative multi-modal object ReID datasets, encompassing diverse spectral modalities. Experimental results show that the proposed method achieves excellent performance on all datasets and is significantly better than current methods in terms of noise immunity. Our code is available at <https://github.com/wanxixi11/UGG-ReID>.

1 Introduction

Multi-modal data has emerged as a significant trend in artificial intelligence [1–4]. Especially driven by large model technology, more and more applications have begun to utilize multi-modal information for comprehensive analysis [5–7]. Multi-modal object Re-IDentification (ReID) [8–11], as a cutting-edge direction of this research, not only broadens the application boundaries of traditional object ReID [12–14], but also effectively makes up for some limitations in cross-modal object ReID [15–17].

Recently, researchers have extensively explored multi-modal feature fusion and matching strategies to bridge the representation gap between different modalities for object ReID [9, 18–20]. For example, Zheng *et al.* [9] propose a progressive fusion method to achieve effective fusion of multi-modal data.

*Corresponding Author

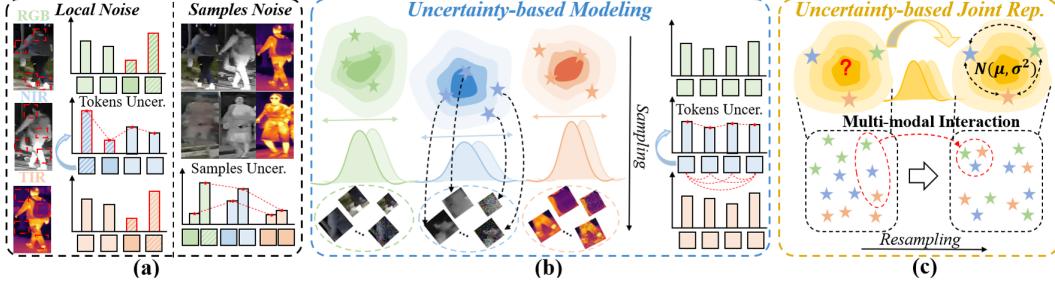


Figure 1: Modeling aleatoric uncertainty in multi-modal ReID. (a) The challenges in multi-modal object ReID. (b) Local uncertainty modeling for modality-specific refinement. (c) Joint uncertainty modeling in samples and modalities for improved fusion.

Wang *et al.* [18] propose HTT that exploits the relationship on unseen test data between heterogeneous modalities to improve performance. On the other hand, existing methods [21–24] have begun to pay attention to the effect of modal noise on the discriminative properties of local regions, such as the introduction of local region alignment, noise suppression module or sample reconstruction, to enhance the robustness and improve the performance of the model. Among these, Zhang *et al.* [22] propose EDITOR to suppress interference from background information and promote feature learn. Wang *et al.* [23] propose utilizing CDA to focus on localized regions of the discriminatory. We can observe that some methods usually assume that the quality and representation of each modal data are balanced and stable, and ignore the local and sample noise disturbances within the modality caused by factors such as occlusion, low resolution [9, 18]. Although existing works [22, 23, 25–27] have achieved some success in mitigating the impact of noise on network performance, there are still significant shortcomings in their approaches when confronted with inconsistent local noise patterns and samples with different noise intensities. As depicted in Fig. 1 (a), the lack of fine-grained local noise learn and sample-level noise handling compromises the robustness of multi-modal fusion, which increases multi-modal uncertainty and impairs overall model performance.

Therefore, to solve the above problems, we provide a robust approach named Uncertainty-Guided Graph model for multi-modal object ReID (UGG-ReID). The proposed method quantifies local aleatoric uncertainty and models their structural dependencies. Following this, leveraging per-sample uncertainty guides feature refinement, promoting more reliable modality interaction. To be specific, we first propose the Gaussian Patch-Graph Representation (GPGR), which encourages fine-grained local features to conform to Gaussian distributions. Meanwhile, a Gaussian patch graph is constructed to explicitly model the dependencies among these local features, thereby capturing fine-grained consistency information. As illustrated in Fig. 1 (b), the uncertainties of local tokens are effectively evaluated and explored via uncertainty-based modeling. Second, we design an Uncertainty-Guided Mixture of Experts (UGMoE) strategy, which dynamically assigns samples to different experts based on their uncertainty levels. Besides, this strategy incorporates an uncertainty-guided routing mechanism to enhance multi-modal interaction. As shown in Fig. 1 (c), multi-modal data is jointly learned to improve the overall performance based on uncertainty. The entire framework is trained in an end-to-end manner. Through extensive experimentation on five public multi-modal object ReID datasets, our method not only achieves competitive performance but also consistently outperforms prior methods in noisy scenarios, validating its effectiveness and robustness.

In summary, the key contributions of this method are outlined as follows:

- We propose the Uncertainty-Guided Graph model for multi-modal object ReID (UGG-ReID). UGG-ReID first enhances features within each modality by considering the distribution of global and local cues and then leveraging experts’ interaction to jointly capture complementary information among multi-modal data, improving model robustness and performance.
- We design a Gaussian Patch-Graph Representation (GPGR) to quantify aleatoric uncertainties for global and local features while modeling their relationships. GPGR can further alleviate the impact of noisy data and effectively reinforce modal-specific information. To our knowledge, this is the first work that leverages uncertainty to **quantify fine-grained local details** and **explicitly model their dependencies** in multi-modal data.
- We introduce the Uncertainty-Guided Mixture of Experts (UGMoE) strategy, which makes different samples select experts based on the uncertainty and also utilizes an uncertainty-

guided routing mechanism to strengthen the interaction between multi-modal features, effectively promoting modal collaboration.

2 Related Works

2.1 Multi-Modal Object ReID

The existing multi-modal ReID methods can be summarized into two types: one focuses on feature fusion and integration between modalities, aiming to alleviate the semantic representation differences between different modalities [19, 24, 28, 29]. For instance, Yang *et al.* [29] propose the tri-interaction enhancement network (TIENet). This method applies spatial-frequency interaction to enhance feature extraction and multi-modal fusion. Wang *et al.* [19] propose a novel method called MambaPro, which utilizes mamba aggregation to fuse the information of multi-modal Object ReID. Zhang *et al.* [24] propose PromptMA to establish effective connections among different multi-modal information. The other type focuses on the local noise interference within the modality and improves the overall recognition performance by enhancing the local region discrimination [10, 21–23]. Zheng *et al.* [10] proposes CCNet, which seeks to simultaneously mitigate identification uncertainty due to modal variability and sample appearance changes by jointly modeling multi-modal heterogeneity and intraclass perturbations under view angle and lighting changes. Zhang *et al.* [22] propose EDITOR that uses Spatial-Frequency Token Selection (SFTS) module to select diverse tokens and suppress the effects of background interference. Wang *et al.* [23] propose the Inverted text with cooperative DEformable Aggregation (IDEA) framework to solve noise interference, enhancing feature robustness. However, the existing methods generally lack explicit modeling of fine-grained local information quality and sample-level uncertainty, which makes it difficult for the model to effectively perceive and suppress local noise in the face of complex modal degradation or multi-source noise interference, which limits its robustness and generalization ability.

2.2 Uncertainty in Multi-Modal Learning

Multi-modal learning faces the challenge of uncertainty from the data layer and the model layer [1, 30–35]. The former is embodied in aleatoric uncertainty, which arises from indelible perceived noise; The latter manifests as an epistemic uncertainty due to the limitations of the model’s capabilities. In practice, aleatoric uncertainty is more common and has a direct impact on model performance. To this end, many works in recent years have focused on the introduction of uncertainty mechanisms to multi-modal learning to effectively identify and suppress unreliable information, improving the performance of the model. Ji *et al.* [30] propose the Probability Distribution Encoder (PDE) module to aggregate all modalities into a probability distribution, framing the uncertainty and optimizing multi-modal representations. Gao *et al.* [1] propose quantifying the intrinsic aleatoric uncertainty of single modality to enhance multi-modal features. Zhang *et al.* [34] propose UMLMC that employs uncertainty-guided meta-learning to mitigate feature-level bias. Therefore, reasonable modeling and quantification of the uncertainty in multi-modal data can not only improve the robustness and generalization ability of multi-modal models but also provide more reliable decision support for practical applications.

3 Methodology

We propose a novel Uncertainty-Guided multi-modal object ReID (UGG-ReID) framework, as shown in Fig. 2. Specifically, the proposed UGG-ReID first designs the Gaussian Patch-Graph Representation (GPGR) model as Fig. 2 (a), which quantifies both global and local uncertainties and models the dependencies between them. This enables richer feature representations for each modality and yields semantically more consistent embeddings. Subsequently, the Uncertainty-Guided Mixture of Experts (UGMoE) strategy makes samples select experts of low uncertainty, promoting multi-modal interactions while mitigating the propagation of excessive noise. Overall, this framework injects a controlled amount of sample noise during the learning process and effectively captures multi-modal information, thereby enhancing the model’s robustness and performance. Below, we will introduce each component of UGG-ReID with details.

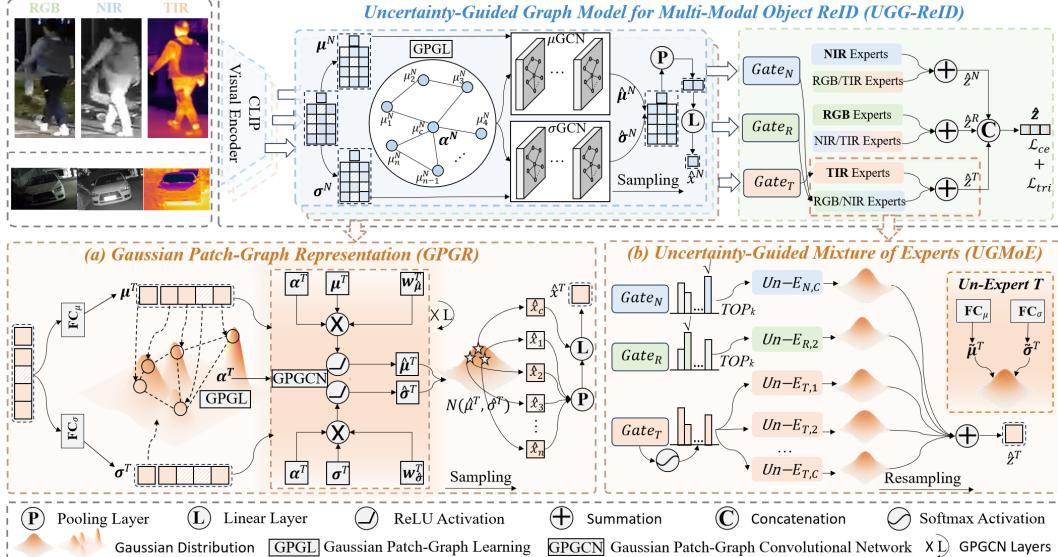


Figure 2: The overall framework of the proposed Uncertainty-Guided multi-modal object ReID (UGG-ReID), which is composed of two main components: Gaussian Patch-Graph Representation (GPGR) and Uncertainty-Guided Mixture of Experts (UGMoE).

3.1 Feature Initialization

To align with prior research [36, 19], we adopt the visual encoder of CLIP with a shared backbone to extract initial features $x^m = \{x_c^m, x_p^m\}$ of multi-modal data. Here, $m \in \{R, N, T\}$ indexes the RGB, Near-Infrared (NIR), and Thermal Infrared (TIR) modalities, while x_c^m and $x_p^m = \{x_1^m, x_2^m \dots x_n^m\}$ denote class tokens and local tokens, respectively. n is the number of local tokens.

3.2 Gaussian Patch-Graph Representation

The proposed Gaussian Patch-Graph Representation (GPGR) model aims to adopt Gaussian distributions as node representations and then conduct message passing between Gaussian distributions. GPGR mainly contains two key components: Gaussian Patch-Graph Learning for relationship modeling and Gaussian Patch-Graph Convolutional Network for message passing. Next, we will introduce the above two modules in detail.

3.2.1 Gaussian Patch-Graph Learning

In this section, we construct a Gaussian Patch-Graph Learning (GPGL) based on the extracted initial features to capture context-aware dependencies. Let $G(V^m, E^m)$ be a Gaussian patch-graph, V^m denotes the node set of the m -th modality and E^m represents the corresponding edge set. Different from deterministic learning, we assume that node representation follows a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$ to distinguish between regions with sufficient and insufficient cues and enhance the ability to learn complex localizations. To be specific, given the initial global-local feature set $\mathbf{x}^m = \{x_c^m, x_1^m, x_2^m \dots x_n^m\} \in \mathbb{R}^{N \times D}$, where $N = n + 1$ and D are the number of nodes and the dimension of each node. We first obtain the mean and standard deviation for global and local nodes as follows,

$$\begin{aligned} \mu_c^m &= \text{FC}_\mu(x_c^m), \quad \mu_i^m = \text{FC}_\mu(x_i^m), \quad i = 1 \dots n, \\ \sigma_c^m &= \text{FC}_\sigma(x_c^m), \quad \sigma_i^m = \text{FC}_\sigma(x_i^m), \quad i = 1 \dots n, \end{aligned} \quad (1)$$

where FC_μ and FC_σ are projection layers to learn the mean and standard deviation, respectively. We denote the means and variances by $\mu^m = \{\mu_c^m, \mu_1^m, \mu_2^m \dots \mu_n^m\} \in \mathbb{R}^{N \times D}$ and $\sigma^m = \{\sigma_c^m, \sigma_1^m, \sigma_2^m \dots \sigma_n^m\} \in \mathbb{R}^{N \times D}$, respectively. Then, we calculate the edge weight in E^m and construct the structural relationship to achieve global and local information interaction. Let α^m be the adjacency matrix and encode structural relationships of m -th modality. Specifically, we compute the similarity by utilizing the mean vectors μ^m to avoid the interference of high uncertain nodes,

$$\alpha_{ij}^m = \text{Similarity}(\mu_i^m, \mu_j^m), \quad (2)$$

where μ_i^m and μ_j^m are the mean representations of i -th and j -th nodes in m -th modality. Similarity is a metric function and we adopt Euclidean distance [37] in our experiments.

3.2.2 Gaussian Patch-Graph Convolutional Network

After contructing the Gaussian Patch-Graph, we further employ a Gaussian Patch-Graph Convolutional Network (GPGCN) to facilitate message passing among distributional nodes. Since each node is modeled as a Gaussian distribution, traditional graph convolution operations are not directly applicable. Thus, we adopt GPGCN to separately conduct message passing based on the mean and variance [38] and thus enable effective propagation of both semantic representations and uncertainty information. Formally, the message passing process is defined as,

$$\begin{aligned}\hat{\boldsymbol{\mu}}^{(m,l+1)} &= \text{ReLU} \left[(\mathcal{D}^m)^{-\frac{1}{2}} \boldsymbol{\alpha}^m (\mathcal{D}^m)^{-\frac{1}{2}} \hat{\boldsymbol{\mu}}^{(m,l)} \mathbf{w}_{\hat{\boldsymbol{\mu}}}^{(m,l)} \right], \\ \hat{\boldsymbol{\sigma}}^{(m,l+1)} &= \text{ReLU} \left[(\mathcal{D}^m)^{-\frac{1}{2}} \boldsymbol{\alpha}^m (\mathcal{D}^m)^{-\frac{1}{2}} \hat{\boldsymbol{\sigma}}^{(m,l)} \mathbf{w}_{\hat{\boldsymbol{\sigma}}}^{(m,l)} \right],\end{aligned}\quad (3)$$

where $\hat{\boldsymbol{\mu}}^{(m,0)} = \boldsymbol{\mu}^m$ and $\hat{\boldsymbol{\sigma}}^{(m,0)} = \boldsymbol{\sigma}^m$. $l = 0, 1 \dots L - 1$ denotes the l -th layers of GGCN and \mathcal{D}^m is corresponding degree matrix of $\boldsymbol{\alpha}^m$. $\mathbf{w}_{\hat{\boldsymbol{\mu}}}^{(m,l)}$ and $\mathbf{w}_{\hat{\boldsymbol{\sigma}}}^{(m,l)}$ are trainable weight matrices. After stacking multiple layers, we can obtain the final output $\hat{\boldsymbol{\mu}}^{(m,L)}$ and $\hat{\boldsymbol{\sigma}}^{(m,L)}$. For convenience, let $\hat{\boldsymbol{\mu}}^m = \hat{\boldsymbol{\mu}}^{(m,L)}$ and $\hat{\boldsymbol{\sigma}}^m = \hat{\boldsymbol{\sigma}}^{(m,L)}$ in the following text. Considering that node representations follow Gaussian distributions, we adopt the sampling operation to obtain final output, i.e., $\hat{\mathbf{x}}^m \sim \mathcal{N}(\hat{\boldsymbol{\mu}}^m, \hat{\boldsymbol{\sigma}}^{2m})$. However, this sampling process is not differentiable. Therefore, we employ the reparameterization trick to generate node representations as follows,

$$\hat{\mathbf{x}}^m = \hat{\boldsymbol{\mu}}^m + \epsilon \hat{\boldsymbol{\sigma}}^m, \quad \epsilon \sim \mathcal{N}(0, I) \quad (4)$$

where ϵ follows a normal distribution. $\hat{\mathbf{x}}^m = \{\hat{x}_c^m, \hat{x}_1^m \dots \hat{x}_n^m\}$ represents the final node representations. To alleviate the effect of extremely unbalanced standard deviation, we adopt a sigmoid function to map $\hat{\boldsymbol{\sigma}}$ to $[0, 1]$. In addition, we also introduce a learnable hyperparameter ϕ that can dynamically adjust the constraint strength of the standard deviation according to the data. This allows the model to flexibly cope with different noise levels while maintaining robustness. Finally, to extract a more effective embedding of global information, local cues are aggregated into global messages to enhance features as follows,

$$\tilde{\mathbf{x}}^m = [\hat{x}_c^m, \eta[\hat{x}_1^m, \hat{x}_2^m \dots \hat{x}_n^m]] W^m, \quad (5)$$

where η denotes a pooling operation. $[\cdot, \cdot]$ is the concatenation operation and W^m is the transformation matrix. To further promote joint aggregation representation, Kullback-Leibler divergence [39] is applied to the class token as,

$$\mathcal{L}_c^m = KL[\mathcal{N}(x_c^m | \mu_c^m, (\sigma_c^m)^2) \| \mathcal{N}(\epsilon | 0, I)] = -\frac{1}{2} (1 + \log(\sigma_c^m)^2 - (\mu_c^m)^2 - (\sigma_c^m)^2). \quad (6)$$

3.3 Uncertainty-Guided Mixture of Experts

Based on the proposed GPGR, we obtain a more abundant feature representation in each modality. To further suppress the possible noise interference in each modality and enhance the multi-modal semantic consistency, we propose the Uncertainty-Guided Mixture of Experts (UGMoE) strategy, which aims to achieve robust and deep multi-modal collaboration by modeling the sample-level uncertainty and guiding the sharing of more expressive experts among modalities.

3.3.1 Uncertainty-Guided Experts Network

In the traditional MoE, the selection of experts is often based on the input content. In the application of multi-modal data, experts are usually selected independently in each modality, which makes it easy to ignore the sample noise and complementarity between modalities [36, 40]. The proposed experts network can bridge this weakness. To be specific, we first model the uncertainty of the extracted features of each modality to enhance the ability of experts to process samples with different noise levels. The multivariate Gaussian distribution that maps the features of each modality is defined as,

$$p(z^m | \tilde{\mathbf{x}}^m) \sim \mathcal{N}(\tilde{\boldsymbol{\mu}}^m, (\tilde{\boldsymbol{\sigma}}^m)^2), \quad (7)$$

where the means $\tilde{\boldsymbol{\mu}}$ and variances $\tilde{\boldsymbol{\sigma}}$ are obtained by two independent fully connected layers, respectively. Then, to model the uncertainty, we resample from the distribution as follows,

$$z^m = \tilde{\boldsymbol{\mu}}^m + \epsilon \tilde{\boldsymbol{\sigma}}^m, \quad \epsilon \sim \mathcal{N}(0, I). \quad (8)$$

However, to ensure the stability and robustness of the model expression, we do not directly use the noisy sampled feature z^m in the prediction process of the final task, but use the mean as the final feature representation of the sample for the downstream decision modeling. In addition to this, to further the uncertainty modeling capabilities, we introduce the KL divergence [39] regularization term to ensure that the sample distribution is close to the normal distribution as follows,

$$\mathcal{L}_s^m = KL[\mathcal{N}(z^m|\tilde{\mu}^m, (\tilde{\sigma}^m)^2) \parallel \mathcal{N}(\epsilon|0, I)] = -\frac{1}{2}(1 + \log(\tilde{\sigma}^m)^2 - (\tilde{\mu}^m)^2 - (\tilde{\sigma}^m)^2). \quad (9)$$

3.3.2 Uncertainty-Guided Routing

To better manage input features effectively, we design an uncertainty-guided routing that includes a gate mechanism to gain modal interaction by selecting different modal experts. The routing process first applies a linear transformation to the input feature. Then, using a softmax activation acts on the result of this transformation to obtain a probability score $S(\tilde{x}^m) \in \mathbb{R}^C$ where C denotes the number of experts. Finally, the TOP_k operation is employed to select the top k ($k=1$) excellent experts of the current modality to learn other modalities, optimizing multi-modal interactions. Thus, each gate involves the total number $C+M-1$ experts, where M denotes the total number of multi-modal data. Besides, to improve the experts' ability, we further add this constraint term [31] as follows,

$$\mathcal{L}_r^m = \frac{1}{C+M-1} \sum_{c=1}^{C+M-1} (\tilde{\sigma}_c^m)^2 S_c^m(\tilde{x}^m). \quad (10)$$

As $\tilde{\sigma}_c^m$ increases, the corresponding expert assigns smaller weights by this constraint.

3.3.3 Interactive Aggregation

We use gate scores as weights to fuse the expert output results by the uncertainty-guided routing operation above as follows,

$$\hat{z}^m = \sum_{c=1}^{C+M-1} S_c(\tilde{x}^m) E_c(\tilde{x}^m). \quad (11)$$

The learned feature tends to be obtained by specific experts, which means that the existence of some experts can not be optimized. To solve this problem, we further add regular terms [31] as,

$$\mathcal{L}_e^m = \frac{1}{C+M-1} \sum_{c=1}^{C+M-1} \left(\frac{1}{B} \sum_{\tilde{X}^m \in B} \mathbb{1} \left\{ \arg \max S_c^m(\tilde{X}^m) = c \right\} \right) \left(\frac{1}{B} \sum_{\tilde{X}^m \in B} S_c^m(\tilde{X}^m) \right), \quad (12)$$

where B denotes the batch size and \tilde{X}^m is the features collection of samples in batch for the m -th modality. The former item refers to the proportion of samples assigned to expert c , and the latter item refers to the proportion of weights assigned by the router to expert c . Then, we aggregate interactive features via the concatenation operation as $\hat{\mathbf{z}} = [\hat{z}^R, \hat{z}^N, \hat{z}^T]$.

3.4 Train Loss

In the training of the proposed UGG-ReID, we combine multiple loss functions to optimize the overall framework. First, $\mathcal{L}_{c,s}^m$ is used to constrain the global features, which represents the sum of \mathcal{L}_c^m and \mathcal{L}_s^m . Then, to improve the network of uncertainty-guided experts, we introduce \mathcal{L}_r^m , which prioritizes experts with low uncertainty by dynamically adjusting expert selection, while adopting \mathcal{L}_e^m to prevent over-reliance on certain experts. Finally, cross-entropy loss \mathcal{L}_{ce} and triplet loss \mathcal{L}_{tri} are used to supervise the entire network. This optimization loss can be expressed by the following formula,

$$\mathcal{L} = \mathcal{L}_{ce} + \mathcal{L}_{tri} + \sum_{m \in \{R, N, T\}} (\lambda_1 \mathcal{L}_{c,s}^m + \lambda_2 \mathcal{L}_r^m + \lambda_3 \mathcal{L}_e^m), \quad (13)$$

where λ_1, λ_2 and λ_3 are the balancing coefficients of this overall loss term.

4 Experiments

In this section, we evaluate the effectiveness of the proposed UGG-ReID on five commonly used datasets and compare it with some state-of-the-art methods.

Table 1: Comparison with state-of-the-art methods on the multi-modal person ReID datasets(in %).

	Methods	Publication	Structure	RGBNT201				Market1501-MM			
				mAP	R-1	R-5	R-10	mAP	R-1	R-5	R-10
Multi-modal	HAMNet [8]	AAAI20	CNN	27.7	26.3	41.5	51.7	60.0	82.8	92.5	95.0
	PFNet [9]	AAAI21	CNN	38.5	38.9	52.0	58.4	60.9	83.6	92.8	95.5
	IEEE [41]	AAAI22	CNN	46.4	47.1	58.5	64.2	64.3	83.9	93.0	95.7
	TIENet [29]	TNNLS25	CNN	54.5	54.4	66.3	71.1	67.4	86.1	94.1	96.0
	UniCat [42]	NIPSW23	ViT	57.0	55.7	-	-	-	-	-	-
	EDITOR [22]	CVPR24	ViT	66.7	68.7	82.2	87.9	77.4	90.8	96.8	98.3
	RSCNet [43]	TCSV24	ViT	68.2	72.5	-	-	-	-	-	-
	HTT [18]	AAAI24	ViT	71.1	73.4	83.1	87.3	67.2	81.5	95.8	97.8
	TOP-ReID [21]	AAAI24	ViT	72.2	75.2	84.9	89.4	82.0	92.4	97.6	98.6
	ICPL-ReID [28]	TMM25	CLIP	75.1	77.4	84.2	87.9	-	-	-	-
	PromptMA [24]	TIP25	CLIP	78.4	80.9	87.0	88.9	83.6	93.3	-	-
	MambaPro [19]	AAAI25	CLIP	78.9	83.4	89.8	91.9	84.1	92.8	97.7	98.7
	DeMo [36]	AAAI25	CLIP	79.7	81.8	89.4	92.5	83.6	93.1	97.5	98.7
	IDEA [23]	CVPR25	CLIP	80.2	82.1	90.0	93.3	-	-	-	-
UGG-ReID		Ours	CLIP	81.2	86.8	92.0	94.7	85.4	94.3	98.4	99.1

4.1 Experiments Setting

Datasets. We conduct five multi-modal object ReID datasets, including two person ReID datasets (e.g., RGBNT201[9], Market1501-MM [41]) and three vehicle datasets (e.g., MSVR310 [10], RGBNT100 [8], WMVEID863 [11]). These datasets pose multi-dimensional challenges such as perspective changes and environmental disturbances, reflecting the broad applicability of this method.

Implementation Details. All experiments are conducted using PyTorch on a single NVIDIA RTX 4090 GPU. A pre-trained CLIP model serves as the visual encoder. Person and vehicle images are resized to 256×128 and 128×256, respectively. We extract features using a 16×16 patching strategy, yielding 128 local tokens and one global token, which jointly serve as nodes in a Gaussian patch-graph for modeling. The model is fine-tuned with Adam (learning rate: 0.00035) for 40 epochs. More details of the experiments are provided in the supplementary material.

Evaluation Protocols. We utilize mAP and Rank to evaluate the performance of the model, where mAP means the accuracy of ReID, while Rank shows the probability that the correct match is included in the top results. The combination of the two can more accurately reflect the ability to identify.

4.2 Comparison with State-of-the-Art Methods

Evaluation on Multi-modal Person ReID. We evaluate our proposed UGG-ReID on two multi-modal person ReID datasets in Table 1. We can observe that both datasets achieve state-of-the-art results for all metrics. Particularly, RGBNT201 [9] outperforms the next most popular method in the metric rank-1 by 3.4%. In addition, our method surpasses DeMo [36] in multiple metrics, highlighting the effectiveness of our architectural design. DeMo [36] relies on modality decoupling to preserve specific cues and uses attention to assign expert weights in MoE. In contrast, our proposed GPGR builds stronger modality-specific representations by incorporating aleatoric uncertainty from local details. Meanwhile, the UGMoE strategy utilizes sample uncertainty to guide expert selection and applies a novel routing strategy to facilitate more effective multi-modal collaboration.

Evaluation on Multi-modal Vehicle ReID. We further conduct experiments on three multi-modal vehicle datasets, which contain challenges such as large view discrepancies and intense glare conditions, to fully validate the robustness and effectiveness of the proposed UGG-ReID. As shown in Table 2, our method maintains stable performance under various complex interference conditions, clearly demonstrating the practical effectiveness of the proposed method in enhancing the model’s generalization capability. Notably, on WMVEID863 [11] with severe dazzle interference, our method outperforms the suboptimal method by 2.8% and 3.6% in the mAP and R-1, respectively. This result further validates the robustness and effectiveness in the face of significant noise interference.

4.3 Ablation Study

To analyze the contribution of each module, we conduct systematic ablation experiments around the two core components, UGMoE and GPGR, on RGBNT201 [9] and WMVEID863 [11]. We first

Table 2: Comparison with state-of-the-art methods on the multi-modal vehicle ReID datasets(in %).

Methods	Publication	Structure	MSVR310		RGBNT100		WMVEID863				
			mAP	R-1	R-5	R-10	mAP	R-1	R-5	R-10	
Multi-modal	HAMNet [8]	AAAI20	CNN	27.1	42.3	74.5	93.3	45.6	48.5	63.1	68.8
	PFNet [9]	AAAI21	CNN	23.5	37.4	68.1	94.1	50.1	55.9	68.7	75.1
	IEEE [41]	AAAI22	CNN	21.0	41.0	61.3	87.8	45.9	48.6	64.3	67.9
	CCNet [10]	INFFUS23	CNN	36.4	55.2	77.2	96.3	50.3	52.7	69.6	75.1
	EDITOR [22]	CVPR24	ViT	39.0	49.3	82.1	96.4	65.6	73.8	80.0	82.3
	RSCNet [43]	TCSV24	ViT	39.5	49.6	82.3	96.6	-	-	-	-
	TOP-ReID [21]	AAAI24	ViT	35.9	44.6	81.2	96.4	67.7	75.3	80.8	83.5
	FACENeT [11]	INFFUS25	ViT	36.2	54.1	81.5	96.9	69.8	77.0	81.0	84.2
	PromptMA [24]	TIP25	CLIP	55.2	64.5	85.3	97.4	-	-	-	-
	MambaPro [19]	AAAI25	CLIP	47.0	56.5	83.9	94.7	69.5	76.9	80.6	83.8
	DeMo [36]	AAAI25	CLIP	49.2	59.8	86.2	97.6	68.8	77.2	81.5	83.8
	IDEA [23]	CVPR25	CLIP	47.0	62.4	87.2	96.5	-	-	-	-
	ICPL-ReID [28]	TMM25	CLIP	56.9	77.7	87.0	98.6	67.2	74.0	81.3	85.6
UGG-ReID		Ours	CLIP	60.1	78.0	88.0	98.1	72.6	80.8	84.2	87.2

Table 3: Ablation study results on the RGBNT201 and WMVEID863 datasets (in %).

UGMoE		GPGR		RGBNT201				WMVEID863			
MoE	Uncer.	PGR	Uncer.	mAP	R-1	R-5	R-10	mAP	R-1	R-5	R-10
(a)	×	×	×	72.2	72.7	82.2	87.3	66.4	72.4	79.4	82.9
(b)	✓	×	×	75.0	76.3	86.5	89.5	68.7	75.8	80.2	82.9
(c)	✓	✓	×	76.0	80.4	88.8	91.7	69.7	77.0	81.5	84.5
(d)	✓	✓	✓	77.3	81.7	90.3	92.1	70.2	77.4	83.5	86.7
(e)	✓	✓	✓	81.2	86.8	92.0	94.7	72.6	80.8	84.2	87.2

use the pre-trained shared CLIP visual encoder as the baseline in Table 3, and gradually introduce each component for comparative analysis. **UGMoE**. Introducing the traditional Mixture of Experts (MoE) strategy to the baseline can bring some performance improvement in Table 3 (b). We further incorporate uncertainty modeling into the MoE strategy and the model performance continues to improve. Compared with the baseline, the proposed method improves mAP/R-1 by 4.8%/7.7% and 3.3%/4.6% on RGBNT201 [9] and WMVEID863 [11], respectively, which indicates that the introduction of uncertainty helps to estimate the expert’s credibility more accurately, thus realizing a more reasonable sample assignment and improving the effectiveness of multi-modal information fusion. **GPGR**. In the case of relying only on the above strategy for multi-modal modeling, the modal-specific information is still not fully explored. To further promote the model’s ability, we introduce GPGR to enhance the modeling ability of fine-grained local cues by using uncertainty. In the experiment, we first evaluate the effect of the standard Patch-Graph Representation (PGR) module in Table 3 (d), and then integrate the uncertainty into the network. The results show that GPGR delivers significant improvements based on the integrated UGMoE. Compared to using UGMoE alone, our method improves mAP/R-1 by 5.2%/6.4% on RGBNT201 [9] and by 2.9%/3.8% on WMVEID863 [11], respectively. This fully shows that GPGR can effectively suppress noise interference and mine richer information. In summary, our proposed method enhances features within each modality and jointly captures complementary information among multi-modal data, improving model robustness and effectiveness. More ablation is provided in the supplementary material.

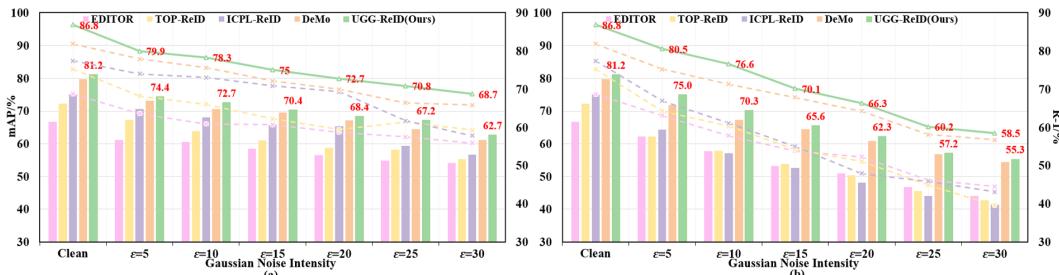


Figure 3: Robustness analysis on RGBNT201. Evaluation results with (a) different levels of Gaussian noise added during dataset generation, and (b) varying noise intensities added during testing after training on clean data.

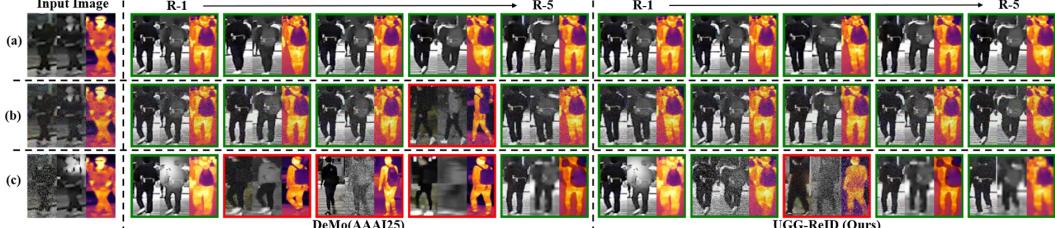


Figure 4: Retrieval results under different testing conditions after training on clean data. (a) Clean. (b) Gaussian noise. (c) Arbitrary noise. Green/Red boxes indicate correct/incorrect retrieval results.

Table 4: Expert-balanced analysis of

Table 5: Efficiency analysis and accuracy comparison with
each modality on WMVEID863. SOTA methods on RGBNT201.

M	Score	E_1	E_2	E_3	E_4
R	Uncer.	0.31	0.27	0.30	0.12
	Gate	0.07	0.12	0.07	0.74
N	Uncer.	0.24	0.24	0.27	0.24
	Gate	0.28	0.26	0.18	0.28
T	Uncer.	0.28	0.26	0.29	0.18
	Gate	0.17	0.21	0.14	0.48

RGBNT201	Params(M)	FLOPs(G)	FPS	mAP	R-1
TOP-ReID [21]	324.5	35.5	398.9	72.2	75.2
EDITOR [22]	119.3	40.8	335.1	66.7	68.7
PromptMA [24]	107.9	36.2	343.5	78.4	80.9
MambaPro [19]	74.8	52.4	243.2	78.9	83.4
DeMo [36]	98.8	35.1	403.6	79.7	81.8
IDEA [23]	91.7	43.7	299.5	80.2	82.1
UGG-ReID	103.2	35.0	371.4	81.2	86.8

4.4 Robustness Analysis

To systematically evaluate the robustness of the proposed UGG-ReID under noise interference, we inject Gaussian noise of different intensities into the RGBNT201 dataset [9], and increase the noise intensity ε from 5 to 30 to generate multiple noisy versions of the dataset. Compared to the four mainstream methods (EDITOR [22], TOP-ReID [21], ICPL-ReID [28] and DeMo [36]), the proposed method maintains superior performance under all noise intensities, as shown in Fig. 3 (a). These results indicate that the method has good robustness and generalization ability.

Furthermore, we evaluate the model’s performance by injecting Gaussian noise of different intensities in the testing phase after completing training on clean data. As Fig. 3 (b) shows, the proposed method is stable and superior to other methods under multi-level noise conditions. Meanwhile, we perform rank-list retrieval evaluations on clean data and different types of noise in Fig. 4. The results show that the proposed method achieves excellent performance under various interference conditions, which is significantly better than the advanced method DeMo [36], which fully reflects its robustness and practicability under complex perceptual interference.

4.5 Expert Balanced Analysis

We show the router average gate scores and uncertainty scores of all test samples for each modality on WMVEID863 in Table 4. We observe that N and T modalities show a more balanced expert activation distribution, while Expert 4 in the R modality has a heavy weight when the uncertainty is low. These results indicate that the routing mechanism dynamically adjusts expert allocation according to different modalities and sample uncertainty.

4.6 Efficiency Analysis

To further validate the effectiveness of our proposed UGG-ReID, we conduct an efficiency analysis, as shown in Table 5. We evaluate the inference speed of each method on the RGBNT201 dataset, using Frames Per Second (FPS) as the evaluation metric. As shown in the Table 5, UGG-ReID reaches an inference speed of 371.4 FPS while maintaining a relatively low parameter count of 103.2 million and a computational cost of 35.0G FLOPs. This speed is only slightly lower than that of the lighter DeMo [36], and is significantly higher than most mainstream approaches, including MambaPro [19] and EDITOR [22]. Notably, despite its high efficiency, UGG-ReID still achieves excellent results.

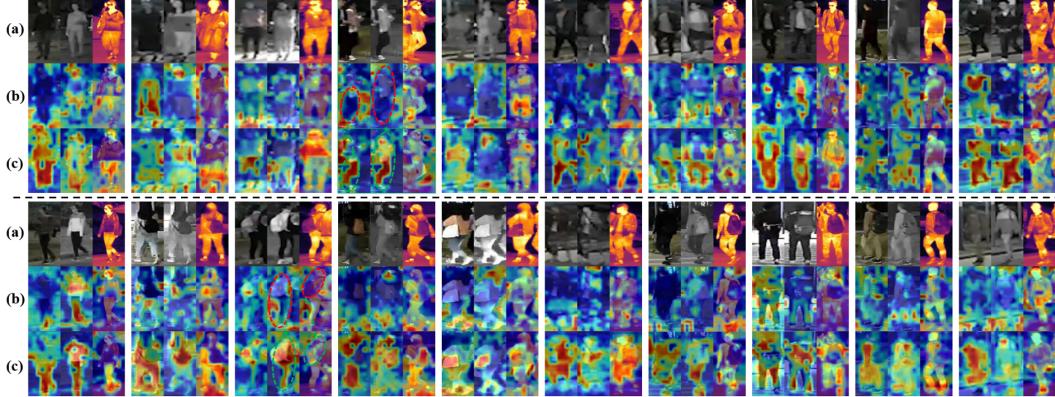


Figure 5: Visualization results of the (a) Input image. (b) Baseline. (c) UGG-ReID (Ours).

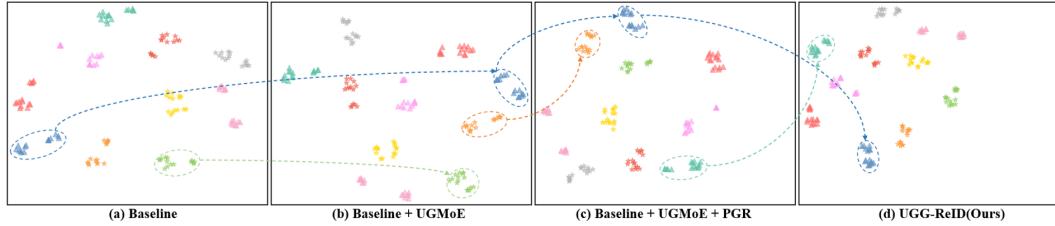


Figure 6: T-SNE visualization of extracted features with different model component combinations on the RGBNT201 dataset.

4.7 Visual Results

Multi-modal Activation Maps: We adopt the Grad-CAM method [44] to visualize the extracted features of each modality. Samples with backpacks are above the dotted line, and samples without backpacks are below the dotted line. The proposed UGG-ReID consistently attends to more semantically rich object regions compared to the baseline model that lacks our module, as shown in Fig. 5. Notably, under challenging conditions such as motion blur or occlusion, our method still effectively focuses on key target features, further validating its robustness and discriminative power. This suggests that our method effectively captures the diversity of multi-modal information through uncertainty-guided multi-modal local and sample-level joint learning, which significantly improves the robustness and the performance of object re-identification.

Multi-modal Feature Distributions: We adopt the T-SNE method [45] to visualize the extracted features to intuitively show the feature distribution of the model under different combinations of modules. With the gradual introduction of the proposed modules, the feature distribution gradually shows obvious clustering in Fig. 6. This compact clustering structure is consistent with the performance enhancement results in Table 3, which further validates the effectiveness of our method in enhancing modal feature representation and multi-modal interaction modeling. More visualizations are provided in the supplementary material.

5 Conclusion

In this paper, we propose a novel Uncertainty-Guided Graph model for multi-modal object (UGG-ReID), effectively mining modal-specific information and boosting modal collaboration for multi-modal object ReID. This UGG-ReID consists of two main aspects. One design is the Gaussian Patch-Graph Representation (GPGR) model for quantifying the aleatoric uncertainty of local and global features and modeling the relationship between them. This manner mitigates the noise interference and enhances the learning ability and robustness of modality-specific information. The other introduces the Uncertainty-Guided Mixture of Experts (UGMoE) strategy to select appropriate experts based on sample uncertainty and facilitate deep multi-modal interactions through a novel routing mechanism. These two core schemes synergistically optimize the overall network structure and achieve the current optimal performance on all five multi-modal object ReID datasets.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (62372003, 62576004), the Natural Science Foundation of Anhui Province (2308085Y40, 2408085J037), the Key Technologies R&D Program of Anhui Province (202423k09020039), and the Open Research Project of the Anhui Provincial Key Laboratory of Security Artificial Intelligence (SAI202401).

References

- [1] Z. Gao, X. Jiang, X. Xu, F. Shen, Y. Li, and H. T. Shen, “Embracing unimodal aleatoric uncertainty for robust multimodal fusion,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 26 866–26 875.
- [2] X. Wang, S. Wang, C. Tang, L. Zhu, B. Jiang, Y. Tian, and J. Tang, “Event stream-based visual object tracking: A high-resolution benchmark dataset and a novel baseline,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 19 248–19 257.
- [3] J. Li, S. Wang, Q. Zhang, S. Yu, and F. Chen, “Generating with fairness: A modality-diffused counterfactual framework for incomplete multimodal recommendations,” in *Proceedings of the ACM on Web Conference 2025*, 2025, pp. 2787–2798.
- [4] L. Bao, X. Zhou, B. Zheng, R. Cong, H. Yin, J. Zhang, and C. Yan, “Ifenet: Interaction, fusion, and enhancement network for v-d-t salient object detection,” *IEEE Transactions on Image Processing*, vol. 34, pp. 483–494, 2025.
- [5] B. McKinzie, Z. Gan, J.-P. Fauconnier, S. Dodge, B. Zhang, P. Dufter, D. Shah, X. Du, F. Peng, A. Belyi, et al., “Mm1: methods, analysis and insights from multimodal llm pre-training,” in *Proceedings of the European Conference on Computer Vision*, 2024, pp. 304–323.
- [6] X. Wang, B. Zhuang, and Q. Wu, “Modaverse: Efficiently transforming modalities with llms,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 26 596–26 606.
- [7] S. Wu, H. Fei, L. Qu, W. Ji, and T.-S. Chua, “Next-gpt: any-to-any multimodal llm,” in *Proceedings of the International Conference on Machine Learning*, 2024.
- [8] H. Li, C. Li, X. Zhu, A. Zheng, and B. Luo, “Multi-spectral vehicle re-identification: A challenge,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, pp. 11 345–11 353.
- [9] A. Zheng, Z. Wang, Z.-H. Chen, C. Li, and J. Tang, “Robust multi-modality person re-identification,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, pp. 3529–3537.
- [10] A. Zheng, X. Zhu, Z. Ma, C. Li, J. Tang, and J. Ma, “Cross-directional consistency network with adaptive layer normalization for multi-spectral vehicle re-identification and a high-quality benchmark,” *Information Fusion*, vol. 100, p. 101901, 2023.
- [11] A. Zheng, Z. Ma, Y. Sun, Z. Wang, C. Li, and J. Tang, “Flare-aware cross-modal enhancement network for multi-spectral vehicle re-identification,” *Information Fusion*, vol. 116, p. 102800, 2025.
- [12] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang, “Omni-scale feature learning for person re-identification,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3701–3711.
- [13] P. Wang, X. Zheng, L. Qing, B. Li, F. Su, Z. Zhao, and H. Chen, “Drformer: A discriminable and reliable feature transformer for person re-identification,” *IEEE Transactions on Information Forensics and Security*, vol. 20, pp. 980–995, 2025.
- [14] A. Zheng, J. Liu, Z. Wang, L. Huang, C. Li, and B. Yin, “Visible-infrared person re-identification via specific and shared representations learning,” *Visual Intelligence*, vol. 1, no. 29, 2023.
- [15] P. Dai, R. Ji, H. Wang, Q. Wu, and Y. Huang, “Cross-modality person re-identification with generative adversarial training,” in *International Joint Conference on Artificial Intelligence*, 2018, pp. 677–683.
- [16] Y. Hao, N. Wang, X. Gao, J. Li, and X. Wang, “Dual-alignment feature embedding for cross-modality person re-identification,” in *Proceedings of the ACM International Conference on Multimedia*, 2019, pp. 57–65.
- [17] S. He, H. Luo, P. Wang, F. Wang, H. Li, and W. Jiang, “Transreid: Transformer-based object re-identification,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 013–15 022.

- [18] Z. Wang, H. Huang, A. Zheng, and R. He, “Heterogeneous test-time training for multi-modal person re-identification,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, pp. 5850–5858.
- [19] Y. Wang, X. Liu, T. Yan, Y. Liu, A. Zheng, P. Zhang, and H. Lu, “Mambapro: Multi-modal object re-identification with mamba aggregation and synergistic prompt,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.
- [20] X. Yang, W. Dong, D. Cheng, N. Wang, and X. Gao, “Tienet: A tri-interaction enhancement network for multimodal person reidentification,” *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–12, 2025.
- [21] Y. Wang, X. Liu, P. Zhang, H. Lu, Z. Tu, and H. Lu, “Top-reid: Multi-spectral object re-identification with token permutation,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023, pp. 5758–5766.
- [22] P. Zhang, Y. Wang, Y. Liu, Z. Tu, and H. 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*, 2024, pp. 17117–17126.
- [23] Y. Wang, Y. Lv, P. Zhang, and H. Lu, “Idea: Inverted text with cooperative deformable aggregation for multi-modal object re-identification,” *arXiv preprint arXiv:2503.10324*, 2025.
- [24] S. Zhang, W. Luo, D. Cheng, Y. Xing, G. Liang, P. Wang, and Y. Zhang, “Prompt-based modality alignment for effective multi-modal object re-identification,” *IEEE Transactions on Image Processing*, vol. 34, pp. 2450–2462, 2025.
- [25] M. Yang, Z. Huang, P. Hu, T. Li, J. Lv, and X. Peng, “Learning with twin noisy labels for visible-infrared person re-identification,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 14288–14297.
- [26] M. Yang, Z. Huang, and X. Peng, “Robust object re-identification with coupled noisy labels,” *International Journal of Computer Vision*, vol. 132, no. 7, pp. 2511–2529, 2024.
- [27] Y. Qin, Y. Chen, D. Peng, X. Peng, J. T. Zhou, and P. Hu, “Noisy-correspondence learning for text-to-image person re-identification,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 27187–27196.
- [28] S. Li, A. Zheng, C. Li, J. Tang, and B. Luo, “Icpl-reid: Identity-conditional prompt learning for multi-spectral object re-identification,” *IEEE Transactions on Multimedia*, 2025.
- [29] X. Yang, W. Dong, D. Cheng, N. Wang, and X. Gao, “Tienet: A tri-interaction enhancement network for multimodal person reidentification,” *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–12, 2025.
- [30] Y. Ji, J. Wang, Y. Gong, L. Zhang, Y. Zhu, H. Wang, J. Zhang, T. Sakai, and Y. Yang, “Map: Multimodal uncertainty-aware vision-language pre-training model,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 23262–23271.
- [31] Z. Gao, D. Hu, X. Jiang, H. Lu, H. T. Shen, and X. Xu, “Enhanced experts with uncertainty-aware routing for multimodal sentiment analysis,” in *Proceedings of the ACM International Conference on Multimedia*, 2024, pp. 9650–9659.
- [32] S. Li, X. Xu, C. He, F. Shen, Y. Yang, and H. Tao Shen, “Cross-modal uncertainty modeling with diffusion-based refinement for text-based person retrieval,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 35, no. 3, pp. 2881–2893, 2025.
- [33] Z. Gao, X. Jiang, H. Chen, Y. Li, Y. Yang, and X. Xu, “Uncertainty-debiased multimodal fusion: Learning deterministic joint representation for multimodal sentiment analysis,” in *Proceedings of the IEEE International Conference on Multimedia and Expo*, 2024, pp. 1–6.
- [34] D. Zhang, M. A. Bashar, and R. Nayak, “A novel multi-modal fusion method based on uncertainty-guided meta-learning,” *Pattern Recognition*, vol. 158, p. 110993, 2025.
- [35] Y. Deng, Z. Chen, C. Li, and J. Tang, “Uncertainty-aware coarse-to-fine alignment for text-image person retrieval,” *Visual Intelligence*, vol. 3, no. 6, 2025.
- [36] Y. Wang, Y. Liu, A. Zheng, and P. Zhang, “Demo: Decoupled feature-based mixture of experts for multi-modal object re-identification,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

- [37] Z. Liu, H. Li, R. Li, Y. Zeng, and J. Ma, “Graph embedding based on euclidean distance matrix and its applications,” in *Proceedings of the ACM International Conference on Information & Knowledge Management*, 2021, pp. 1140–1149.
- [38] D. Zhu, Z. Zhang, P. Cui, and W. Zhu, “Robust graph convolutional networks against adversarial attacks,” in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, p. 1399–1407.
- [39] J. Chang, Z. Lan, C. Cheng, and Y. Wei, “Data uncertainty learning in face recognition,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5709–5718.
- [40] S. Yun, I. Choi, J. Peng, Y. Wu, J. Bao, Q. Zhang, J. Xin, Q. Long, and T. Chen, “Flex-moe: Modeling arbitrary modality combination via the flexible mixture-of-experts,” in *Proceedings of the Conference on Neural Information Processing System*, 2024.
- [41] Z. Wang, C. Li, A. Zheng, R. He, and J. Tang, “Interact, embed, and enlarge: Boosting modality-specific representations for multi-modal person re-identification,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022, pp. 2633–2641.
- [42] J. Crawford, H. Yin, L. McDermott, and D. Cummings, “Unicat: Crafting a stronger fusion baseline for multimodal re-identification,” *arXiv preprint arXiv:2310.18812*, 2023.
- [43] Z. Yu, Z. Huang, M. Hou, J. Pei, Y. Yan, Y. Liu, and D. Sun, “Representation selective coupling via token sparsification for multi-spectral object re-identification,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 35, no. 4, pp. 3633–3648, 2025.
- [44] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-cam: Visual explanations from deep networks via gradient-based localization,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2017, pp. 618–626.
- [45] D. Kobak and G. C. Linderman, “Initialization is critical for preserving global data structure in both t-sne and umap,” *Nature Biotechnology*, vol. 39, pp. 156–157, 2021.
- [46] P. Kaushik, A. Kortylewski, and A. Yuille, “A bayesian approach to ood robustness in image classification,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 22 988–22 997.
- [47] S. Li, X. Xu, Y. Yang, F. Shen, Y. Mo, Y. Li, and H. T. Shen, “Dcel: Deep cross-modal evidential learning for text-based person retrieval,” pp. 6292–6300, 2023.
- [48] Q. Zha, X. Liu, Y.-m. Cheung, X. Xu, N. Wang, and J. Cao, “Ugncl: Uncertainty-guided noisy correspondence learning for efficient cross-modal matching,” in *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2024, pp. 852–861.

A Supplementary Material

In the supplementary materials, we provide a detailed description of the UGMoE strategy and present additional experiments to further validate the robustness and effectiveness of the proposed UGG-ReID.

A.1 Details of UGMoE

In the main text, we set the value of k in Top k to be 1. In this case, each modality will have $C - 1$ unique experts and M shared experts, so the total number of experts is $C + M - 1$. If $k \neq 1$, we analyze it further to get $C + k(M - 1)$ experts for each modality. Thus, Eq. 10 is defined as follows,

$$\mathcal{L}_r^m = \frac{1}{C + k(M - 1)} \sum_{c=1}^{C+k(M-1)} (\tilde{\sigma}_c^m)^2 S_c^m(\tilde{x}^m). \quad (14)$$

Next, we utilize gate scores as weights to fuse the expert output results by the above routing operation as follows,

$$\hat{z}^m = \sum_{c=1}^{C+k(M-1)} S_c(\tilde{x}^m) E_c(\tilde{x}^m). \quad (15)$$

The learned feature tends to be obtained by specific experts, which means that the existence of some experts can not be optimized. To solve this problem, we further add regular terms [31] as,

$$\mathcal{L}_e^m = \frac{1}{C + k(M - 1)} \sum_{c=1}^{C+k(M-1)} \left(\frac{1}{B} \sum_{\tilde{X}^m \in B} 1 \left\{ \arg \max S_c^m(\tilde{X}^m) = c \right\} \right) \left(\frac{1}{B} \sum_{\tilde{X}^m \in B} S_c^m(\tilde{X}^m) \right), \quad (16)$$

where B denotes the batch size and \tilde{X}^m is the features collection of samples in batch for the m -th modality. The former item refers to the proportion of samples assigned to expert c , and the latter item refers to the proportion of weights assigned by the router to Expert c . B denotes the batch size. Finally, we aggregate interactive features via the concatenation operation as $\hat{\mathbf{z}} = [\hat{z}^R, \hat{z}^N, \hat{z}^T]$.

A.2 Comparison with Prior Works

For conceptual comparison, we first utilize a Gaussian-based random graph for object representation, where nodes are described by Gaussian distributions to represent the uncertainty of image patches in the presence of noise, whereas previous works generally employ a deterministic graph model, represented by a feature vector. We design a Gaussian Patch-Graph Representation (GPGR) to quantify aleatoric uncertainties for global and local features while modeling their relationships. To our knowledge, this work is the first attempt to exploit a random patch-graph model for the object ReID problem. Second, for the Mixture-of-Experts approach, we design the Uncertainty-Guided Mixture of Experts (UGMoE) strategy, which enables different samples to select experts based on uncertainty and utilizes an uncertainty-guided routing mechanism to strengthen the interaction between multi-modal features, effectively promoting modal collaboration.

For a technical comparison, we first further analyze the impact of different uncertainty modeling approaches on the performance of multi-modal object ReID [1, 30, 31]. As shown in Table 6, works EUAR [31] and EAU [1] are able to perceive inter-sample uncertainty. In contrast, MAP [30] introduces a more comprehensive uncertainty modeling mechanism to quantify the uncertainty

Table 6: Component-wise comparison of different methods on RGBNT201 (in %).

Method	Uncer.	MoE	Local	Gloabl	Graph	mAP	R-1
EUAR [31]	✓	✓	✗	✓	✗	74.1	77.6
EAU [1]	✓	✗	✗	✓	✗	75.6	80.3
MAP [30]	✓	✗	✓	✓	✗	76.8	78.2
DeMo [36]	✗	✓	✓	✓	✗	79.7	81.8
UGG-ReID	✓	✓	✓	✓	✓	81.2	86.8

of local cues. Although the above methods take uncertainty into account, they either neglect the modeling of uncertainty in local cues or the structural relationships between local regions. Second, we perform comparisons under different MoE strategies. As shown in Table 6, we substitute two existing MoE methods [31, 36]. Compared with DeMo [36], UGMOE better exploits the diversity of samples by introducing uncertainty modeling, and compared with EUAR [31], UGMOE further strengthens the interaction between different modalities.

A.3 Details of Experiments

A.3.1 Experiments Setting

Datasets. To comprehensively evaluate the generalization ability of the proposed UGG-ReID framework, we conduct experiments on five public datasets. These include two person re-identification datasets, RGBNT201 [9] and Market1501-MM [41], as well as three challenging vehicle re-identification benchmarks: MSVR310 [10], RGBNT100 [8], and WMVEID863 [11]. These datasets collectively reflect a wide range of real-world scenarios and associated challenges. Table 7 summarizes the partition protocols and the specific challenges posed by each dataset.

Table 7: Details of the datasets partition settings and their corresponding challenges, /* represents ID/Sample.

	RGBNT201	Market1501-MM	MSVR310	RGBNT100	WMVEID863
Train	171/3951	751/12936	155/1032	50/8675	603/10446
Query	30/836	750/3368	52/591	50/1715	210/2904
Gallery	30/836	751/15913	155/1055	50/8575	272/3678
Challenges	Wide Views, Occlusions	Simulate the Night Scene	Longer Time Span, Complex Conditions	Different Views, Illumination Issue	Intense Flare

Implementation Details. For all experiments, we set the number of experts at $C = 4$ and utilize $k = 1$ for the TOP_k selection. The loss terms are weighted with $\lambda_1 = 0.1$ and $\lambda_2, \lambda_3 = 0.0001$, respectively. The number of layers for GPGCN L is set to 2. Our code is implemented in Python using the PyTorch framework and will be released publicly upon acceptance.

A.3.2 Ablation Analysis

To verify the role of each loss in the model, we conduct systematic ablation experiments, as shown in Table 8. $\mathcal{L}_{c,s}$ represents the sum of \mathcal{L}_c and \mathcal{L}_s , which is used to impose constraints on the global token. From the experimental results, one can observe that when the $\mathcal{L}_{c,s}$ constraint on the global token is removed, the performance of the model on multiple evaluation indicators decreases, indicating that the constraint has a positive effect on improving modeling ability. Then, \mathcal{L}_r^m is removed to verify performance for adding the loss constraint on the expert. We can find that adding the loss, our mAP/R-1 increases by 2.2%/3.1% and 2.0%/1.6% in RGBNT201 [9] and WMVEID863 [11], respectively, which verifies its enhancement effect on the expert selection strategy. Finally, we verify the effectiveness of \mathcal{L}_e^m , which aims to ensure that the number of similar samples assigned to each expert in the training process is balanced. Meanwhile, the expert weights are relatively evenly distributed among the experts, and the experimental results show that it can effectively prevent the imbalance of distribution among experts and improve the ability of the model.

Table 8: Ablation results for different loss on the RGBNT201 and WMVEID863 datasets (in %).

Loss	Type	RGBNT201				WMVEID863			
		mAP	R-1	R-5	R-10	mAP	R-1	R-5	R-10
(a)	w/o $\mathcal{L}_{c,s}$	78.8	84.8	89.6	91.7	69.6	76.9	82.4	85.8
(b)	w/o \mathcal{L}_r^m	79.0	83.7	91.6	94.1	70.6	79.2	84.2	86.8
(c)	w/o \mathcal{L}_e^m	81.0	84.6	90.7	92.5	71.2	78.3	84.5	87.7
(d)	UGG-ReID	81.2	86.8	92.0	94.7	72.6	80.8	84.2	87.2

Table 9: Results of the analysis on hyperparameters C and k on the RGBNT201 and WMVEID863 datasets (in %).

Expert C	RGBNT201		WMVEID863	
	mAP	R-1	mAP	R-1
2	80.3	83.7	71.4	78.8
3	80.8	85.2	72.0	80.1
4	81.2	86.8	72.6	80.8
5	80.0	84.6	72.0	79.7
6	80.1	83.7	71.1	78.1

Top k	RGBNT201		WMVEID863	
	mAP	R-1	mAP	R-1
0	78.3	83.1	71.3	78.8
1	81.2	86.8	72.6	80.8
2	79.9	84.6	71.7	79.2
3	80.2	85.3	71.2	78.5
4	79.0	81.6	70.3	77.6

Table 10: Results of the analysis on hyperparameters L and n on the RGBNT201 dataset (in %).

Layers L	RGBNT201			
	mAP	R-1	mAP	R-1
1	80.1	85.3	91.3	93.9
2	81.2	86.8	92.0	94.7
3	79.1	84.6	90.8	93.3
4	78.4	82.5	90.3	92.8
5	76.3	80.7	89.1	91.7

Nodes n	RGBNT201			
	mAP	R-1	mAP	R-1
32	79.3	83.7	91.3	93.9
64	80.1	82.9	90.0	94.0
96	81.4	84.6	90.4	92.6
128	81.2	86.8	92.0	94.7
160	79.8	83.6	90.2	91.9

A.3.3 Hyperparameter Analysis

We analyze the effects of the hyperparameters C and k on model performance, where C controls the number of experts and k denotes the number of shared experts selected for each modality. As shown in Table 9, a moderate increase in C enhances the model’s expressive capacity, while an appropriate choice of k strikes a balance between stability and flexibility. This facilitates dynamic collaboration and complementarity among experts, ultimately improving overall model performance.

We further analyze the local nodes n of GPGN and Layers L of GPGCN of the GPGR for the effect of the model in the RGBNT201 dataset in Table 10. For nodes n , we observe that $n=128$ achieves excellent results. Too few nodes are not enough to cover rich local information, and too many introduce redundancy and noise, interfering with graph structure learning. For layers L , GPGCN works best when $L=2$. The number of layers is too shallow and may lead to insufficient fusion of local structures, while too deep may cause over-smoothing, resulting in the loss of discriminative representation of nodes and weakening the expression ability of local discriminative features.

A.3.4 Visual Results

Visualization of Rank List. To analyze the performance of the proposed UGG-ReID method in cross-camera retrieval scenarios, we perform rank-list visualization of the retrieval results of different methods as Fig. 7. Compared with baseline and baseline+UMoE, UGG-ReID can rank the objects more accurately, demonstrating stronger model robustness and discriminative ability.

Visualization of Class Activation Maps. As shown in Fig. 8, we visualize the proposed UGG-ReID using Class Activation Maps (CAMs) [44]. The results further demonstrate that our approach is capable of capturing discriminative local regions, even under complex environmental conditions.

A.4 Discussion

Multi-modal object ReID exploits fine-grained local cues and the complementary information of modalities to effectively enhance the robustness and accuracy of recognition in complex scenarios [9, 21, 23, 24, 36]. As is well known, significant distributional differences exist among different modalities, and noise arising from sample quality and environmental factors further impacts the accuracy of feature representations. The proposed UGG-ReID effectively guides the feature fusion process by explicitly quantifying local and sample-level epistemic uncertainties and modeling the relationship between them, enhancing the model’s robustness and effectiveness. UGG-ReID is the

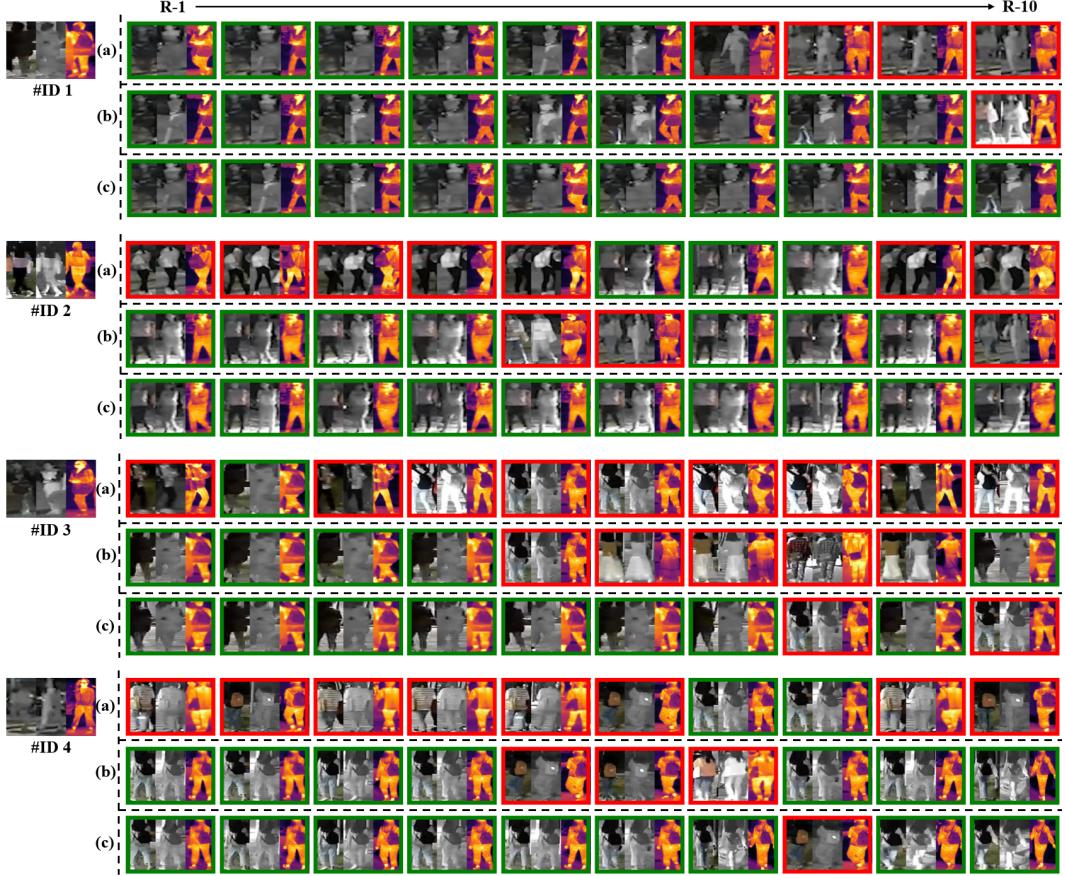


Figure 7: Rank-list visualizations for four persons from the RGBNT201 dataset under different model configurations: (a) Baseline, (b) Baseline + UGMoE, and (c) UGG-ReID (Ours).

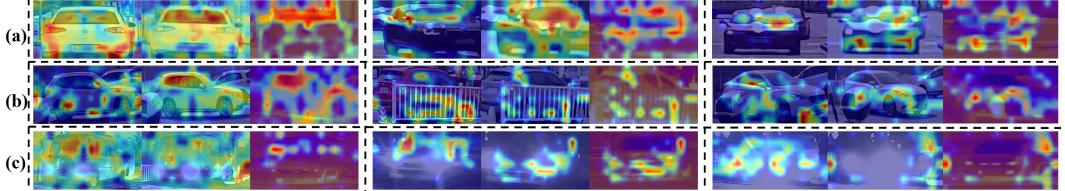


Figure 8: Visualization of Class Activation Maps (CAMs) under different environmental conditions for six vehicles from the WMVEID863 dataset: (a) Normal, (b) Occlusion, and (c) Intense Flare.

first work that leverages uncertainty to quantify fine-grained-local details and explicitly model their dependencies in multi-modal data.

Limitations and In the Future. Our framework employs uncertainty-guided learning to enhance robustness against local noise; it may still struggle under extreme conditions where local cues are heavily corrupted or missing. In future work, we will focus on advancing uncertainty quantification and reasoning techniques, exploring the integration of Bayesian inference and evidence theory into multi-modal object ReID [46–48]. This aims to enhance the model’s robustness to modality and label noise, thereby improving its overall performance and reliability in complex environments.