

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		RGBNT201				WMVEID863			
Type		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 GPGL 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

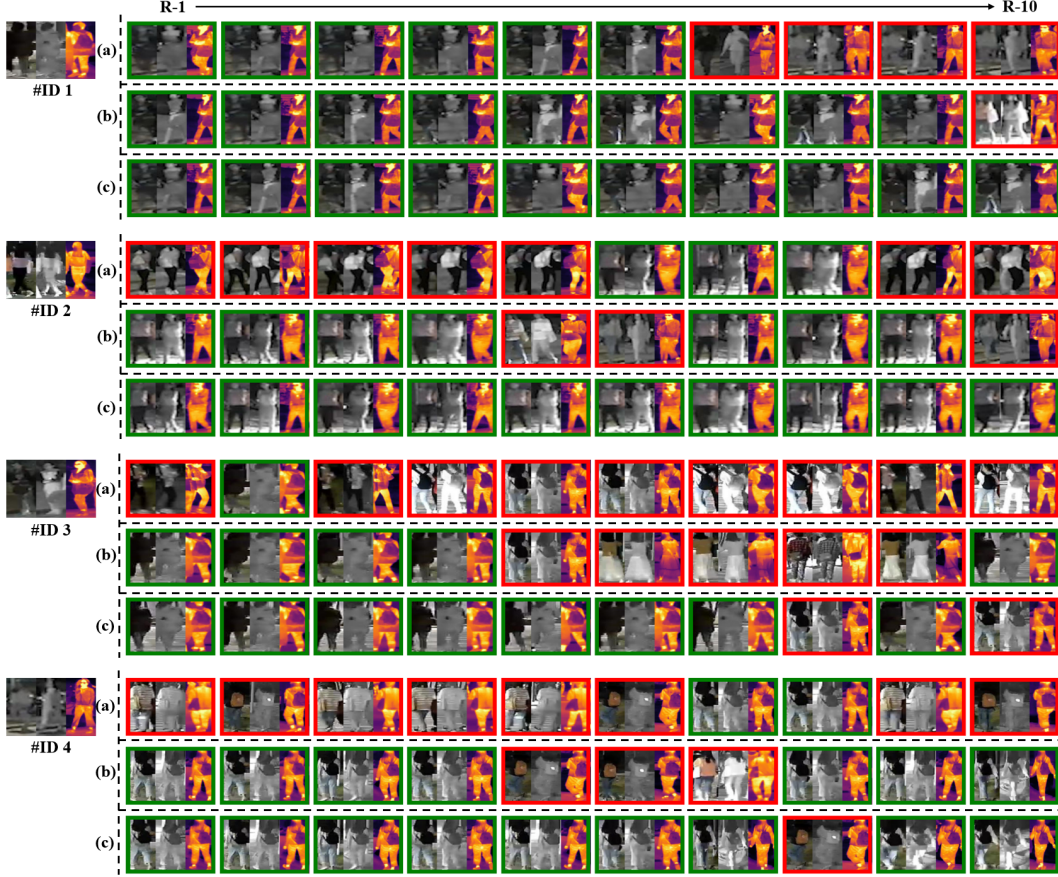


Figure 7: Rank-list visualizations for four persons from the RGBNT201 dataset under different model configurations: (a) Baseline, (b) Baseline + UGMoE, and (c) UGG-ReID (Ours).

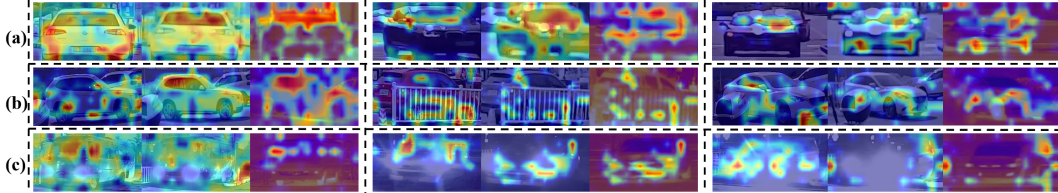


Figure 8: Visualization of Class Activation Maps (CAMs) under different environmental conditions for six vehicles from the WMVEID863 dataset: (a) Normal, (b) Occlusion, and (c) Intense Flare.

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.