MixER: From Cross-Modal to Mixed-Modal Visible-Infrared Re-Identification

Supplementary Material

A. Proofs:

In Section 3, our model is designed to represent input images through two distinct and complementary feature types: modality-related and modality-erased representations. To ensure the independence of these representations, we minimize the mutual information (MI) between the data distributions of them. This is achieved by applying an orthogonal loss to their feature representations. The modality-related features are intended to capture ID-aware information specific to the modality, whereas the modality-erased features are explicitly designed to exclude any modality-specific information. To ensure that the extracted features satisfy these conditions, we use constrained optimization to maximize the mutual information between the joint distribution of modality-related and modality-erased features and the label distribution. In this section, we prove that our proposed loss functions meet these constraints and align with the properties of mutual information.

A.1. Backgrounds

The following highlights the main properties of mutual information:

P 1 (Nonnegativity). *For every pair of random variables X and Y:*

$$MI(X;Y) \geq 0 \tag{A.1}$$

P 2. *For random variables X, Y that are independent:*

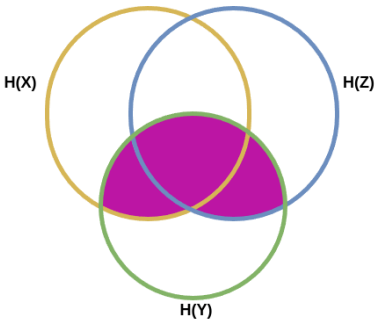
$$MI(X;Y) = 0. \tag{A.2}$$

P 3 (Monotonicity). *For every three random variables X, Y and Z:*

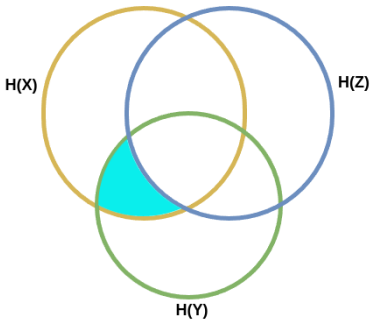
$$MI(X;Y;Z) \leq MI(X;Y) \tag{A.3}$$

P 4. *For every three random variables X, Y and Z, the mutual information of joint distribution X and Z to Y is (Fig. A.1a):*

$$MI(X,Z;Y) = MI(X;Y) + MI(Z;Y) - MI(X;Z;Y) \tag{A.4}$$



(a) MI$(X,Z;Y)$



(b) MI$(X;Y|Z)$

Figure A.1. Venn diagram of theoretic measures for three variables X, Y, and Z, represented by the left, right and bottom circles, respectively.

P 5 (Conditional). *For every three random variables X, Y and Z, the conditional mutual information is(Fig. A.1b):*

$$MI(X;Y|Z) = MI(X;Y) - MI(X;Y;Z) \tag{A.5}$$

Also, we described some hypotheses that are used in our learning:

Hypothesis 1. *Following [12], the orthogonality between $\mathbf{z}_m^r$ and $\mathbf{z}_m^e$ can be regarded as a relaxation of independence:*

$$\forall(\mathbf{z}_m^r, \mathbf{z}_m^e) \sim (Z_m^r, Z_m^e),\ \mathbf{z}_m^r, \mathbf{z}_m^e \perp \mathbf{z}_m^e \ \Rightarrow\ \textit{MI}(Z_m^r; Z_m^e) \simeq 0$$

Hypothesis 2. *Following [1, 12], we posit that if Z is a representation of X, then Z is conditionally independent of all other variables in the system given X. This is formally expressed as:*

$$\forall A, B \quad I(A; Z \mid X, B) = 0. \tag{A.6}$$

Definition 1 (**Sufficiency**). *A representation Z of X is sufficient for Y if and only if:*

$$\textit{MI}(X; Y|Z) = 0 \Longleftrightarrow \textit{MI}(X; Y) = \textit{MI}(Z; Y) \tag{A.7}$$

Any model with access to a sufficiently informative representation Z must be able to predict Y with at least the same level of accuracy as if it had access to the original data X. Representation Z is considered sufficient for Y if and only if the task-relevant information remains unchanged during the encoding process. If the cross-entropy loss between Z and Y is minimized, then, as suggested in [1], Z can be assumed to be a sufficient representation of X for the task Y.

Definition 2 (**Data Processing Inequality**). *Let three random variables form the Markov chain $Y \to X \to Z$, implying that the conditional distribution of Z depends only on X and is conditionally independent of X, we have:*

$$\textit{MI}(X; Y) \geq \textit{MI}(Z; Y). \tag{A.8}$$

The data processing inequality (DPI) is a fundamental inequality in information theory that states the mutual information between two random variables cannot increase through processing.

Theorem 1. *If Z_m^e and Z_m^r are independent, then $\textit{MI}(Z_m^e, Z_m^r; Y) = \textit{MI}(Z_m^e; Y) + \textit{MI}(Z_m^r; Y)$.*

Proof. For independent random variables Z_m^e and Z_m^r, we have the following relationship:

$$\text{MI}(Z_m^e; Z_m^r) = 0.$$

Also, w.r.t to P 4:

$$\text{MI}(Z_m^r, Z_m^e; Y) = \text{MI}(Z_m^r; Y) + \text{MI}(Z_m^e; Y) - \text{MI}(Z_m^r; Z_m^e; Y),$$

and P 3:

$$\text{MI}(Z_m^r; Z_m^e; Y) \leq \text{MI}(Z_m^r; Z_m^e) = 0,$$

so, we have:

$$\text{MI}(Z_m^e, Z_m^r; Y) = \text{MI}(Z_m^e; Y) + \text{MI}(Z_m^r; Y) \tag{A.9}$$

□

A.2. Minimizing $\text{MI}(Z_m^e; M)$

In order to minimizing $\text{MI}(Z_m^e; M)$, we use adversarial learning approach and convert $Z_m^e \in \mathcal{Z}_m^e$ to $Z_m^o \in \mathcal{Z}_m^o$ with deterministic and learnable function $\mathcal{W} : \mathcal{Z}_m^e \to \mathcal{Z}_m^o$:

$$Z_m^o = \mathcal{W}(Z_m^e; \theta_W). \tag{A.10}$$

Then, we maximize the $\text{MI}(Z_m^o; M)$ with a standard SGD approach through the optimization of θ_W and reversed SGD through the model's parameters. Maximizing $\text{MI}(Z_m^o; M)$ is achieved by minimizing cross-entropy loss between the estimated modality label from Z_m^o as $\hat{m}$ and ground truth label, m ($\mathcal{L}_\text{m}$ in main manuscript). This approximation is formulated in **Proposition 1**.

Proposition 1. *Let Z and Y be random variables with domains $\mathcal{Z}$ and $\mathcal{Y}$, respectively. Minimizing the conditional cross-entropy loss of predicted label $\hat{Y}$, denoted by $\mathcal{H}(Y; \hat{Y}|Z)$, is equivalent to maximizing the $\textit{MI}(Z; Y)$*

Proof. Let us define the MI as entropy,

$$\text{MI}(Z, Y) = \underbrace{\mathcal{H}(Y)}_{\delta} - \underbrace{\mathcal{H}(Y|Z)}_{\xi} \tag{A.11}$$

Since the domain $\mathcal{Y}$ does not change, the entropy of the identity δ term is a constant and can therefore be ignored. Maximizing $\text{MI}(Z, Y)$ can only be achieved through a minimization of the ξ term. We show that $\mathcal{H}(Y|Z)$ is upper-bounded by the cross-entropy loss, and minimizing such loss results in minimizing the ξ term. By expanding its relation to the cross-entropy [**?**]:

$$\mathcal{H}(Y; \hat{Y}|Z) = \mathcal{H}(Y|Z) + \underbrace{\mathcal{D}_{\text{KL}}(Y||\hat{Y}|Z)}_{\geq 0}, \tag{A.12}$$

where we have:

$$\mathcal{H}(Y|Z) \leq \mathcal{H}(Y; \hat{Y}|Z), \tag{A.13}$$

where minimizing $\mathcal{H}(Y; \hat{Y}|Z)$ results minimizing $\mathcal{H}(Y|Z)$. □

A.3. Maximizing $\text{MI}(Z^e_m; Y|M)$

Eq. 8 of main manuscript can be rewritten w.r.t P 5 as :

$$\text{MI}(Z^e_m; Y|M) = \text{MI}(Z^e_m; Y) - \text{MI}(Z^e_m; Y; M). \tag{A.14}$$

For the second part RHS:

$$\text{MI}(Z^e_m; Y; M) \leq \text{MI}(Z^e_m; M),$$

where the $\mathcal{L}_{\text{m}}$ in main manuscript is minimized, results $\text{MI}(Z^e_m; M) \simeq 0$ and :

$$\text{MI}(Z^e_m; Y; M) \simeq 0.$$

So, the Eq. A.14 is:

$$\text{MI}(Z^e_m; Y|M) \simeq \text{MI}(Z^e_m; Y). \tag{A.15}$$

To maximize the $\text{MI}(Z^e_m; Y)$, the $\mathcal{L}_{\text{meid}}$ is minimized w.r.t to **Proposition 1**.

A.4. Minimizing $\text{MI}(Z^r_m; Y|M)$

To enforce the modality-related features Z^r_m to leverage identity-aware information that is dependent on modality (i.e., specific identity-discriminative information), we minimize the amount of identity-aware information in these features that disregards the modality. Below, we demonstrate that $\text{MI}(Z^r_m; Y \mid M)$ is upper-bounded by zero if the features Z^r_m are sufficient for both tasks Y (identity) and M (modality) simultaneously. To ensure that the modality-related features Z^r_m serve as sufficient representations of the input images X for both detecting modality and identifying identity, the loss function $\mathcal{L}_{\text{mrid}}$ (as defined in the main manuscript) is applied to these features.

Theorem 2. *If the representation Z^r_m of X is sufficient for both Y and M, then:*

$$\textit{MI}(Z^r_m; Y \mid M) = 0. \tag{A.16}$$

Proof. From the definition of a sufficient representation Z^r_m for a task M (Definition 1), we have:

$$\text{MI}(X; M \mid Z^r_m) = 0 \iff \text{MI}(X; M) = \text{MI}(Z^r_m; M) \Rightarrow \text{MI}(Z^r_m; M \mid X) = 0. \tag{A.17}$$

Similarly, for task Y, we have:

$$\text{MI}(Z^r_m; Y \mid X) = 0. \tag{A.18}$$

Expanding $\text{MI}(Z^r_m; Y \mid X)$, we obtain:

$$\text{MI}(Z^r_m; Y \mid X) = \text{MI}(Z^r_m; Y; M \mid X) + \text{MI}(Z^r_m; Y \mid X, M) = 0. \tag{A.19}$$

For the first term on the RHS of Eq. A.19, we have:

$$\text{MI}(Z^r_m; Y; M \mid X) = \text{MI}(Z^r_m; M \mid X) - \text{MI}(Z^r_m; M \mid X, Y) = 0,$$

where $\text{MI}(Z_m^r; M \mid X) = 0$ follows from Eq. A.18, since Z_m^r is a sufficient representation of X for task M, and $\text{MI}(Z_m^r; M \mid X, Y) = 0$ is due to Hypothesis 2. For the second term on the RHS of Eq. A.19, we have:

$$\text{MI}(Z_m^r; Y \mid X, M) = \text{MI}(Z_m^r; Y \mid M) - \text{MI}(Z_m^r; Y; X \mid M) = 0. \tag{A.20}$$

Since $\text{MI}(Z_m^r; Y; X \mid M) \leq \text{MI}(Z_m^r; Y \mid M)$, and $\text{MI}(Z_m^r; Y \mid M) = \text{MI}(Z_m^r; Y; X \mid M)$, it follows that:

$$\text{MI}(Z_m^r; Y; X \mid M) = 0. \tag{A.21}$$

Thus, $\text{MI}(Z_m^r; Y \mid M) = 0$, completing the proof. □

B. Additional Experiments

B.1. Implementation Details of Open-Source SOTA Methods

In this section, we present implementation details for each open-source state-of-the-art (SOTA) method used in our manuscript. For each method, we use the hyperparameter based on their paper or official code in GitHub:

- **DDAG** [43]: This method employs ResNet-50 as the backbone with stride one at the last layer, with input images resized to 288×144 pixels that are augmented with zero-padding and horizontal flipping. Based on the DDAG paper[43], 8 people with 4 V and 4 I images are selected in the batch, and $p = 3$ is set.
- **DEEN** [45]: A modified ResNet-50 is used as the backbone, enhanced with DEE modules that introduce two additional branches to the network. Input images are resized to 344×144 pixels. During training, augmentations such as Random Erase and Random Channel augmentation are applied. At inference time, the original image and its horizontally flipped counterpart are both processed through the backbone, and the average of the extracted features is used as the final representation. For evaluation under our mixed-modal settings, we removed the flipping process and instead concatenated the features from all DEE branches to create the final representative features.
- **MPANet** [40]: This method also employs ResNet-50 as the backbone, with an additional convolutional layer designed to detect more discriminative regions in the feature space. The final representative feature vector is constructed by concatenating part features and global features obtained from a Global Average Pooling (GAP) layer. Input images are resized to 344×144 pixels, and augmentations such as Random Erase and Random Channel augmentation are applied during training.
- **SGEIL** [12]: This method employs two ResNet-50 backbones, one for visible images and the other for infrared images, along with an additional ResNet-50 backbone for shape images. Input images are resized to 288×144 pixels, with augmentations similar to those used in DEEN. During training, model weights are updated using SGD, while an exponential moving average (EMA) is simultaneously applied to update the backbone weights. At the end of training, the best performance between the SGD and EMA models is selected. Note that SGEIL requires shape images for training, and since this information is available only for the SYSU-MM01 dataset, its performance on RegDB and LLCM is not reported.
- **SAAI** [11]: Similar to MPANet, this method uses ResNet-50 as the backbone, with an additional convolutional layer and learnable parameters to identify part prototypes. The final representative feature vector is obtained by concatenating part features and global features from the GAP layer. Input images are resized to 288×144 pixels, and Random Erase and Random Channel augmentation are applied during training. An Affinity Inference Module is used during inference to rerank the gallery.
- **IDKL** [36]: This method uses ResNet-50 as the backbone, with additional branches added to layers 3 and 4 to extract modality-specific features. Input images are resized to 388×144 pixels, and Random Erase and Random Channel augmentation are applied during training. During inference, k-reciprocal encoding is applied to rerank the gallery, enhancing the retrieval process.

B.2. Additional Mixed-Modal Results

In the main manuscript, we reported the performance of mixed-modal settings for existing datasets with infrared query images and mixed-modal gallery images. Table B.1, presents the performance with visible query. Across all mixed settings in SYSU-MM01, MixER consistently outperforms the compared methods, achieving the highest Rank-1 accuracy (R1) and mean Average Precision (mAP) scores. Specifically, for the Mix setting, our method achieves a notable improvement in mAP (87.29%) compared to the next best method, IDKL (84.78%), showcasing its robustness in handling mixed gallery conditions. In more challenging settings like Mix-Cam and Mix-ID, MixER significantly outperforms the SOTA, demonstrating its ability to adapt to modality and identity constraints effectively.

In the RegDB dataset, which focuses on visible-infrared retrieval, MixER achieves the highest scores across both the Mix and Mix-ID settings. The proposed method achieves a remarkable mAP of 92.78% in the Mix setting, outperforming the best-performing baseline (IDKL) by over 4.4%. Similarly, for the Mix-ID setting, MixER achieves a substantial improvement in mAP (81.42%) compared to SAAI (68.82%). On the LLCM dataset, MixER maintains competitive performance, achieving the highest scores in both the Mix and Mix-ID settings. In particular, MixER improves mAP in the Mix-ID setting to 45.18%, which surpasses the previous best method, DEEN, by a notable margin (43.65%). These results highlight the generalizability and robustness of our method in addressing varying cross-modal and identity constraints.

The proposed MixER consistently achieves superior performance across all datasets and settings, highlighting its effectiveness in extracting discriminative modality-related and modality-erased features. The substantial improvements in challenging settings, such as Mix-ID, underscore the effectiveness of the disentangling strategy employed by MixER, which allows it to handle both modality and identity variations effectively. Unlike competing methods, MixER achieves a balanced improvement in both Rank-1 accuracy and mAP, demonstrating its robustness in retrieval precision and ranking performance.

Method	SYSU-MM01								RegDB				LLCM			
	Mix		**Mix-Cam**		**Mix-Cam-ID**		**Mix-ID**		**Mix**		**Mix-ID**		**Mix**		**Mix-ID**	
	R1	**mAP**	**R1**	**mAP**	**R1**	**mAP**	**R1**	**mAP**	**R1**	**mAP**	**R1**	**mAP**	**R1**	**mAP**	**R1**	**mAP**
DDAG [43]	97.59	79.62	94.68	76.61	92.96	73.42	41.09	45.85	99.9	77.53	45.39	47.13	99.24	51.31	22.50	17.94
MPANet [40]	**97.94**	83.80	95.23	80.98	94.10	78.91	54.54	57.24	**100**	84.32	60.24	61.57	99.17	59.61	25.08	18.82
DEEN [45]	95.79	82.18	92.29	80.32	89.85	77.32	59.13	61.03	99.95	88.49	75.24	71.45	99.25	73.73	57.90	43.65
SGEIL [12]	96.76	80.52	94.05	78.74	91.10	74.82	48.72	52.96	-	-	-	-	-	-	-	-
SAAI [11]	97.63	83.88	95.22	81.50	93.82	79.08	55.08	57.61	100	88.19	69.22	68.82	99.57	70.60	46.12	34.63
IDKL [36]	98.25	84.78	96.15	82.51	94.87	79.85	54.00	57.51	99.95	88.33	71.70	70.64	99.57	70.54	39.25	32.04
MixER (ours)	97.14	**87.29**	**96.27**	**85.67**	**94.95**	**84.79**	**70.96**	**70.76**	**100**	**92.78**	**85.39**	**81.42**	**99.80**	**74.59**	**58.87**	**45.18**

Table B.1. Accuracy of the proposed method and open-sourced state-of-the-art methods on the SYSU-MM01 (single-shot setting), RegDB, and LLCM datasets in different mixed gallery settings. Visible images are chosen as the query.

B.3. Additional Ablation Studies

B.3.1. Computational Complexity

We compare the training times of several state-of-the-art methods in Cross-Modal ReID to demonstrate how much additional computational burden is added due to the augmentation of these methodologies with our method. The training times have been documented in Table B.2. Note that each method has its own variable number of training epochs, keeping in mind the optimal number of epochs suggested for these methods. We argue that the overall increase in computational time is a reasonable trade-off with a projected performance increase, thus highlighting the superiority of our method. The upper-bound model represents a best-case scenario with three separate SAAI models trained independently for visible, infrared, and VI images. Also, to compare with the upper-bound, using MixER is more efficient.

Method	**Training time**	**Training time with MixER**	**# epochs**	**# of param**	**# of params with MixER**	**Flops**	**Flops with MixER**
DDAG	111	137	80	40M	54M	0.5T	0.6T
DEEN	686	732	151	61M	75M	1.3T	1.5T
SGEIL	597	631	120	87M	101M	0.9T	0.97T
SAAI	78	101	160	72M	85M	0.75T	0.8T
IDKL	182	204	180	88M	106M	0.95T	1.1T
MPANet	144	179	140	74.8M	88M	0.9T	1.02T
Upper-Bound	184	-	160	216M	-	2.25T	-

Table B.2. Training times for state-of-the-art Cross-Modal ReID methods (on SYSY-MM01 dataset) compared to the training times of the same methods modified with our method. Time has been reported in minutes.

Table B.3 compares the performance of the baseline SAAI method [11], its modified version enhanced with our proposed MixER framework, and an upper-bound model across mixed-modal, cross-modal, and uni-modal settings on the SYSU-MM01 dataset. The results show that the SAAI+MixER model consistently outperforms the baseline SAAI in mixed-modal settings, achieving notable improvements in both Rank-1 accuracy and mAP. For instance, in the Mix setting, SAAI+MixER achieves an mAP of 80.47% compared to 74.59% for the baseline. In the challenging Mix-ID setting, it improves mAP

from 53.30% to 62.45%, demonstrating the effectiveness of MixER in disentangling modality-related and modality-erased features.

In cross-modal settings, SAAI+MixER also demonstrates robustness, achieving a slight improvement in mAP for the "All" setting (71.08% vs. 69.71%) while maintaining competitive performance in uni-modal scenarios. In the I→I and V→V settings, SAAI+MixER either matches or outperforms the baseline without compromising intra-modality retrieval. Notably, in some cases, SAAI+MixER exceeds the upper-bound, such as in the Mix setting where it achieves an mAP of 80.47

Method	Mixed-Modal								Cross-Modal				Uni-Modal			
	Mix		Mix-Cam		Mix-Cam-ID		Mix-ID		All		Indoor		I→I		V→V	
	R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP
SAAI[11]	96.01	74.59	90.63	72.51	84.30	65.94	52.49	53.30	73.87	69.71	84.19	82.59	89.29	93.06	98.24	93.46
SAAI[11] +MixER	97.27	80.47	92.66	76.81	88.51	73.24	66.23	62.45	74.25	71.08	-	-	92.11	94.57	98.60	94.08
Upper-bound	95.74	78.20	91.18	74.94	85.44	68.71	59.38	55.70	73.87	69.71	84.19	82.59	88.06	92.80	98.85	95.20

Table B.3. Performance of SAAI[11] VI-ReID technique in mixed, cross, and uni-modal settings on the SYSU-MM01 dataset compared to upper-bound. The upper-bound is a model that contains three separate SAAI models for V,I, and VI images.

B.3.2. The influence of hyperparameters.

In this section, we present a bar chart (Fig. B.1) to examine the detailed influence of hyperparameters by gradually increasing their value. As we can see, the best performance is achieved when λ_1 is set to 0.4, λ_2 is set to 0.6, and λ_3 is set to 0.4, respectively. The upward trend of the bars demonstrates the effectiveness of each loss.

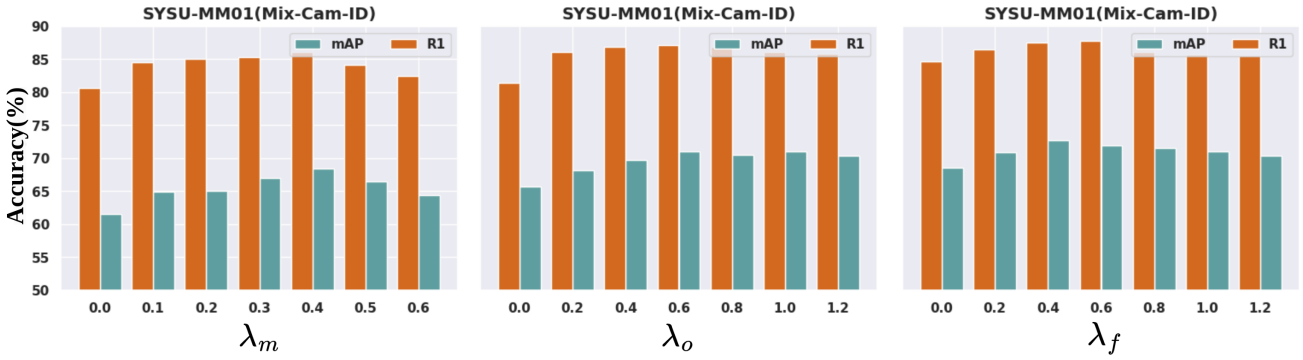


Figure B.1. Influence of different λ_m,λ_o and λ_f values on SYSU-MM01 in Mix-Cam-ID evaluation.

B.3.3. Choice of Backbone.

We test the performance of our method on two main choices of backbones, ResNet [17], and ViT [5]. For vision transformer models, we resized images to 224 by 224 pixels, and we used the extracted feature from the last layer as representative features. We observe that despite its success in recent times [**?**], standard ViT models struggle to perform on par with ResNet. This is likely due to the large image resolution ViTs (224×224) are designed to train on. This size allows patches to be as large as 16×16. However, our input images were originally resized to a shape 288×144 restricts us from using a ViT architecture reliably as the ViT architecture dimensions depend on input dimensions. In this regard, ResNet is a much more flexible backbone that can work on a wide range of image resolutions and therefore delivers better performance. All models were trained following the standard implementation settings as described in Section 4.1 of the main paper, with the exception of ViT inputs being resized to 224×224, instead of 288×144.

Backbone	# Parameters	Cross-Modal		Mix-Cam-ID	
		R1	*mAP*	*R1*	*mAP*
ResNet-18	29.02M	67.68	63.51	80.49	62.61
ResNet-50	58.15M	73.43	70.92	87.56	72.0
ViT-B-16	102.87M	55.47	54.97	72.23	57.51
ViT-L-16	331.55M	52.57	53.04	68.14	57.18

Table B.4. The influence of the choice of baseline backbone on the performance of the proposed method.

B.3.4. Choice of Fusion method.

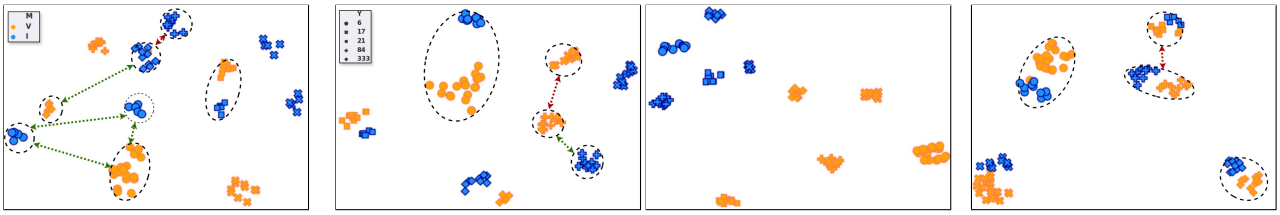
In MixER, the concatenation is used for fusion of $\mathbf{z}_m^e$ and $\mathbf{z}_m^r$. In order to show the effectiveness of Modality-related and modality-erased features in MixER, we test the cross-attention [51] on $\mathbf{z}_m^e$ and $\mathbf{z}_m^r$ before global-average pooling. The results are shown in **??**. Although it increases the mixed settings performance, it increases the number of parameters and latency.

Fusion Method	**# Parameters**	**# Flops**	**Mix-Cam**		**Mix-Cam-ID**	
			R1	***mAP***	***R1***	***mAP***
MixER + Concatention	58.15M	0.8T	91.77	76.35	87.56	72.70
MixER + Cross-Attention	67.33M	0.94T	92.86	77.12	88.09	73.21

Table B.5. The influence of the choice of baseline backbone on the performance of the proposed method.

B.4. Feature Distributions Visualization

We visualized the feature distributions of the baseline and our modules using UMAP [31] in the 2D space, as shown in Fig. B.2(a-d). The results indicate that the modality-erased feature brings embeddings of the same person closer across modalities compared to the baseline (see circular features in Fig. B.2(a,b)). Meanwhile, the modality-related component pushes apart intra-modality features of different individuals. Together, MixER effectively leverages both components to better separate identities and reduce modality discrepancy within the mixed-modal gallery.



(a) Baseline (b) Modality-Erased (c) Modality-Related (d) MixER

Figure B.2. The distribution of feature embeddings in the 2D feature space, where orange and blue colors denote the V and I. The samples with the same shape are from the same person. The green and red arrows show the distance between the same person's features and different persons, respectively.

References

[1] Alessandro Achille and Stefano Soatto. Information dropout: Learning optimal representations through noisy computation. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2897–2905, 2018. 2

[2] Mahdi Alehdaghi, Arthur Josi, Rafael MO Cruz, and Eric Granger. Visible-infrared person re-identification using privileged intermediate information. In *ECCVws*, pages 720–737. Springer, 2022. 1, 3

[3] Mahdi Alehdaghi, Arthur Josi, Pourya Shamsolmoali, Rafael MO Cruz, and Eric Granger. Adaptive generation of privileged intermediate information for visible-infrared person re-identification. *arXiv preprint arXiv:2307.03240*, 2023. 3

[4] Mahdi Alehdaghi, Pourya Shamsolmoali, Rafael MO Cruz, and Eric Granger. Bidirectional multi-step domain generalization for visible-infrared person re-identification. *arXiv preprint arXiv:2403.10782*, 2024. 3

[5] Dosovitskiy Alexey. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv: 2010.11929*, 2020. 6

[6] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *Proceedings of the 35th International Conference on Machine Learning*, pages 531–540. PMLR, 2018. 4

[7] Lei Cai, Zhengyang Wang, Hongyang Gao, Dinggang Shen, and Shuiwang Ji. Deep adversarial learning for multi-modality missing data completion. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1158–1166, 2018. 3

[8] Dapeng Chen, Hongsheng Li, Xihui Liu, Yantao Shen, Jing Shao, Zejian Yuan, and Xiaogang Wang. Improving deep visual representation for person re-identification by global and local image-language association. In *Proceedings of the European conference on computer vision (ECCV)*, pages 54–70, 2018. 3

[9] Seokeon Choi, Sumin Lee, Youngeun Kim, Taekyung Kim, and Changick Kim. Hi-cmd: hierarchical cross-modality disentanglement for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10257–10266, 2020. 3

[10] Zhenyu Cui, Jiahuan Zhou, and Yuxin Peng. Dma: Dual modality-aware alignment for visible-infrared person re-identification. *IEEE Transactions on Information Forensics and Security*, 2024. 1

[11] Xingye Fang, Yang Yang, and Ying Fu. Visible-infrared person re-identification via semantic alignment and affinity inference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11270–11279, 2023. 5, 6, 7, 4

[12] Jiawei Feng, Ancong Wu, and Wei-Shi Zheng. Shape-erased feature learning for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22752–22761, 2023. 1, 2, 3, 4, 6, 7, 5

[13] Yujian Feng, Feng Chen, Guozi Sun, Fei Wu, Yimu Ji, Tianliang Liu, Shangdong Liu, Xiao-Yuan Jing, and Jiebo Luo. Learning multi-granularity representation with transformer for visible-infrared person re-identification. *Pattern Recognition*, 164:111510, 2025. 3

[14] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 214–229. Springer, 2020. 3

[15] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015. 5

[16] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190, 2023. 3

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[19] Weizhen He, Yiheng Deng, Shixiang Tang, Qihao Chen, Qingsong Xie, Yizhou Wang, Lei Bai, Feng Zhu, Rui Zhao, Wanli Ouyang, et al. Instruct-reid: A multi-purpose person re-identification task with instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17521–17531, 2024. 3

[20] Zhipeng Huang, Jiawei Liu, Liang Li, Kecheng Zheng, and Zheng-Jun Zha. Modality-adaptive mixup and invariant decomposition for rgb-infrared person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1034–1042, 2022. 3

[21] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*, pages 5583–5594. PMLR, 2021. 3

[22] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[23] Vladimir V Kniaz, Vladimir A Knyaz, Jirí Hladuvka, Walter G Kropatsch, and Vladimir Mizginov. Thermalgan: Multimodal color-to-thermal image translation for person re-identification in multispectral dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018. 1, 3

[24] Yi-Lun Lee, Yi-Hsuan Tsai, Wei-Chen Chiu, and Chen-Yu Lee. Multimodal prompting with missing modalities for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14943–14952, 2023. 3

[25] He Li, Mang Ye, Ming Zhang, and Bo Du. All in one framework for multimodal re-identification in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17459–17469, 2024. 3

[26] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34: 9694–9705, 2021. 3

[27] Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Foundations and trends in multimodal machine learning: Principles, challenges, and open questions. *arXiv preprint arXiv:2209.03430*, 2022. 3

[28] Wei Liu, Xin Xu, Hua Chang, Xin Yuan, and Zheng Wang. Mix-modality person re-identification: A new and practical paradigm. *ACM Trans. Multimedia Comput. Commun. Appl.*, 2025. 1, 3, 5, 6

[29] Yan Lu, Yue Wu, Bin Liu, Tianzhu Zhang, Baopu Li, Qi Chu, and Nenghai Yu. Cross-modality person re-identification with shared-specific feature transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13379–13389, 2020. 3

[30] Zefeng Lu, Ronghao Lin, and Haifeng Hu. Disentangling modality and posture factors: Memory-attention and orthogonal decomposition for visible-infrared person re-identification. *IEEE Transactions on Neural Networks and Learning Systems*, 2024. 3

[31] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018. 7

[32] Dat Tien Nguyen, Hyung Gil Hong, Ki Wan Kim, and Kang Ryoung Park. Person recognition system based on a combination of body images from visible light and thermal cameras. *Sensors*, 17(3):605, 2017. 5

[33] Yongsheng Pan, Mingxia Liu, Yong Xia, and Dinggang Shen. Disease-image-specific learning for diagnosis-oriented neuroimage synthesis with incomplete multi-modality data. *IEEE transactions on pattern analysis and machine intelligence*, 44(10):6839–6853, 2021. 3

[34] Zhiqi Pang, Lingling Zhao, Yang Liu, Gaurav Sharma, and Chunyu Wang. Inter-modality similarity learning for unsupervised multi-modality person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 3

[35] Hyunjong Park, Sanghoon Lee, Junghyup Lee, and Bumsub Ham. Learning by aligning: Visible-infrared person re-identification using cross-modal correspondences. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12046–12055, 2021. 1, 3

[36] Kaijie Ren and Lei Zhang. Implicit discriminative knowledge learning for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 393–402, 2024. 1, 2, 3, 6, 7, 4, 5

[37] Guan'an Wang, Tianzhu Zhang, Jian Cheng, Si Liu, Yang Yang, and Zengguang Hou. Rgb-infrared cross-modality person re-identification via joint pixel and feature alignment. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019. 1, 3

[38] Hu Wang, Yuanhong Chen, Congbo Ma, Jodie Avery, Louise Hull, and Gustavo Carneiro. Multi-modal learning with missing modality via shared-specific feature modelling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15878–15887, 2023. 3

[39] Ancong Wu, Wei-Shi Zheng, Shaogang Gong, and Jianhuang Lai. Rgb-ir person re-identification by cross-modality similarity preservation. *International journal of computer vision*, 128(6):1765–1785, 2020. 5

[40] Qiong Wu, Pingyang Dai, Jie Chen, Chia-Wen Lin, Yongjian Wu, Feiyue Huang, Bineng Zhong, and Rongrong Ji. Discover cross-modality nuances for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4330–4339, 2021. 3, 4, 5, 6

[41] Yang Yang, Tianzhu Zhang, Jian Cheng, Zengguang Hou, Prayag Tiwari, Hari Mohan Pandey, et al. Cross-modality paired-images generation and augmentation for rgb-infrared person re-identification. *Neural Networks*, 128:294–304, 2020. 3

[42] M. Ye, X. Lan, Z. Wang, and P. C. Yuen. Bi-directional center-constrained top-ranking for visible thermal person re-identification. *IEEE Transactions on Information Forensics and Security*, 15:407–419, 2020. 3

[43] Mang Ye, Jianbing Shen, David J. Crandall, Ling Shao, and Jiebo Luo. Dynamic dual-attentive aggregation learning for visible-infrared person re-identification. In *Computer Vision – ECCV 2020*, pages 229–247, Cham, 2020. Springer International Publishing. 3, 6, 4, 5

[44] 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. *arXiv preprint arXiv:2001.04193*, 2020. 1, 3

[45] Yukang Zhang and Hanzi Wang. Diverse embedding expansion network and low-light cross-modality benchmark for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2153–2162, 2023. 1, 5, 6, 4

[46] Yukang Zhang, Yan Yan, Yang Lu, and Hanzi Wang. Adaptive middle modality alignment learning for visible-infrared person re-identification. *International Journal of Computer Vision*, 133(4):2176–2196, 2025. 3

[47] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pages 1116–1124, 2015. 8

[48] Liang Zheng, Yi Yang, and Alexander G Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016. 3

[49] Ruochen Zheng, Lerenhan Li, Chuchu Han, Changxin Gao, and Nong Sang. Camera style and identity disentangling network for person re-identification. In *BMVC*, page 66, 2019. 3

[50] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, Mingliang Xu, and Yi-Dong Shen. Dual-path convolutional image-text embeddings with instance loss. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 16(2): 1–23, 2020. 3

[51] Xiaoling Zhou, Zetao Jiang, and Idowu Paul Okuwobi. Cafnet: Cross-attention fusion network for infrared and low illumination visible-light image. *Neural Processing Letters*, 55(5):6027–6041, 2023. 7

[52] Xiao Zhou, Yujie Zhong, Zhen Cheng, Fan Liang, and Lin Ma. Adaptive sparse pairwise loss for object re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19691–19701, 2023. 3

[53] Xiaoke Zhu, Minghao Zheng, Xiaopan Chen, Xinyu Zhang, Caihong Yuan, and Fan Zhang. Information disentanglement based cross-modal representation learning for visible-infrared person re-identification. *Multimedia Tools and Applications*, 82(24):37983–38009, 2023. 3

[54] Yuanxin Zhu, Zhao Yang, Li Wang, Sai Zhao, Xiao Hu, and Dapeng Tao. Hetero-center loss for cross-modality person re-identification. *Neurocomputing*, 386:97–109, 2020. 3