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

[1]Mahdi Alehdaghi,[1]Rajarshi Bhattacharya, [2]Pourya Shamsolmoali, [1]Rafael M. O. Cruz, and [1]Eric Granger
[1]LIVIA, Dept. of Systems Engineering, ETS Montreal, Canada
[2]Dept. of Computer Science, University of York, UK
{mahdi.alehdaghi, rajarshi.bhattacharya}.1@ens.etsmtl.ca, pshams55@gmail.com,
{rafael.menelau-cruz, eric.granger}@etsmtl.ca

Abstract

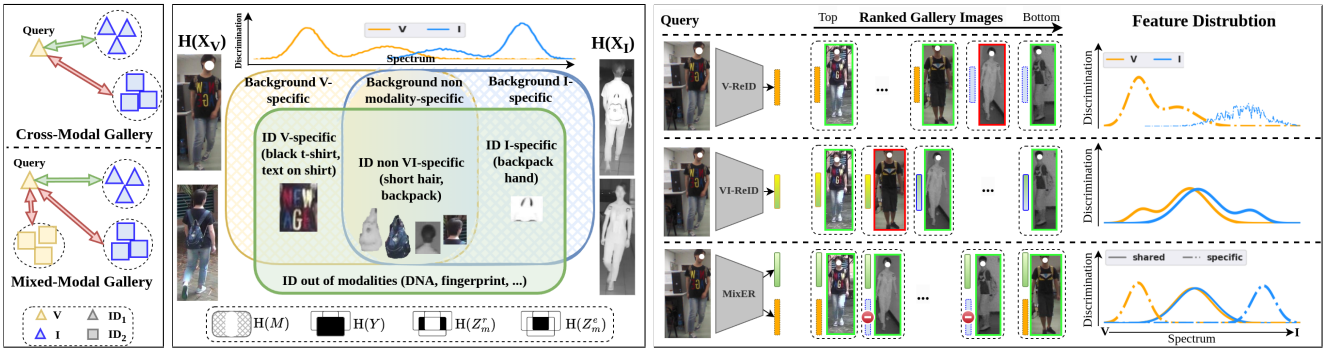
Visible-infrared person re-identification (VI-ReID) aims to match individuals across different camera modalities, a critical task in modern surveillance. While most existing methods focus on cross-modality matching, real-world systems often involve mixed galleries containing both V and I images, where state-of-the-art methods struggle due to large domain shifts and low discrimination across modalities. These challenges arise because same-modality gallery images may have smaller domain gaps but correspond to different identities. To address this, we propose more comprehensive and challenging mixed-modal evaluation settings that better reflect real-world conditions. This paper also introduces the Mixed Modality-Erased and -Related (MixER) method, which disentangles modality-specific and modality-shared identity information through orthogonal decomposition, modality confusion, and ID-modality-related objectives. MixER improves feature robustness across modalities, improving performance in both cross- and mixed-modal settings. Extensive experiments on SYSU-MM01, RegDB, and LLCM show that MixER can achieve state-of-the-art performance with a single backbone and displays strong versatility across diverse mixed-modal scenarios. Our code is available: https://github.com/alehdaghi/MixVI-ReID.

1. Introduction

VI-ReID is crucial for surveillance, matching individuals across cameras and lighting conditions. Existing methods focus on cross-modality matching, where a V or I image is queried against a gallery from the other modality (Fig. 1a, top). However, real-world systems capture both modalities continuously, producing a mixed gallery of V and I images (Fig. 1a, bottom). This setting is more realistic and challenging, as same-modality matches (e.g., V-to-V) have smaller domain gaps than cross-modality matches (e.g., I-to-V). Future VI-ReID methods must handle such mixed galleries by balancing cross- and same-modality matching for reliable real-world performance.

VI-ReID methods are generally expected to perform well in mixed-modality settings, which can be interpreted in various ways. We identify four distinct interpretations based on the identity and camera relationships between query and gallery images. In each setting, specific images are removed to assess the robustness of the matching model. The mixed-modal setting introduced in [25], which considers only same-modality images of the same identity as the query, lacks informativeness and realism. To address this, we propose more realistic and challenging evaluation settings that include both same-modality images of the query identity and different individuals. This design tests model robustness against low-modality gaps, ensuring effective identity discrimination within the same modality while also handling cross-modal variations. However, we show that two specific settings pose significant challenges, where state-of-the-art uni-modal and cross-modal ReID methods struggle to achieve high performance. A cost-effective VI-ReID solution in these settings should rely on a single backbone model to extract discriminative features and dynamically adapt them based on intra-modality or inter-modality matching requirements. Fig. 1b illustrates the identity-discriminative information present in the two modalities, which can be categorized into modality-specific and modality-shared attributes. While V-ReID models primarily extract modality-specific features, VI-ReID methods aim to capture modality-invariant identity-discriminative features for effective cross-modal matching.

Given the substantial discrepancy between I and V modality images, uni-modal models typically struggle to remain discriminative across modalities (as illustrated in the first row of Fig.1c). To address this challenge, VI-ReID methods often focus on minimizing this discrepancy between their extracted feature representations. For example, GAN methods [20, 33] are used to translate the modalities and shared-backbones [1, 8, 10, 31, 32, 40, 41] are used to map images into a modality-invariant feature space, creating a shared representation that minimizes modality dis-



(a) Cross vs Mixed (b) Venn Diagram (c) Mixed-Modal matching

Figure 1. (a) VI-ReID of a query image matched against a cross-modal and mixed-modal gallery. (b) With VI-ReID methods, V, I, and ID information reveal modality-specific and modality-shared ID features. (c) Mixed-modal approaches leverage these features for matching, while uni-modal methods are limited to intra-modality matching due to a large modality gap, and cross-modal methods enable inter-modal matching by extracting shared features. Our MixER method disentangles these features to reduce the gap in shared features and enhance the discrimination in modality-specific features. $H(\cdot)$ measures the entropy of a random variable.

crepancies. Despite the significant improvements in recent years, the features extracted by these methods often contain some modality-specific information, which weakens the effectiveness of inter-modal matching and limits the minimization of modality gap in the feature space.

A limitation of current VI-ReID methods in mixed matching tasks is their lack of attention to identity features that exist only in one modality, as shown in the second row of Fig.1c. Some attributes, such as shirt patterns that are only relevant in the V modality or material textures unique to I, are modality-specific and, therefore, unavailable in the other modality (see Fig.1b). For accurate cross-modal matching, this modality-specific information should be erased from identity representations, only allowing access to modality-shared information across both V and I. However, effectively separating modality-specific from modality-invariant is challenging, as it requires effective unsupervised disentanglement to identify them.

To address the limitations of VI-ReID in mixed-modal scenarios, this paper introduces a **Mix**ed Modality-**E**rased and Modality-**R**elated (MixER) ID-discriminative feature learning approach. By isolating discriminative modality-specific features within one subspace, our approach promotes modality-erased features to capture robust discriminative semantic concepts across modality variations within the other subspace. Implicit Discriminative Knowledge Learning (IDKL) [32] uses modality-specific features to enhance modality-shared ones through knowledge distillation and implicit similarity. In contrast, our approach enforces an orthogonal complement structure. This constraint ensures that modality-related features remain independent of shared features. This allows discovering a diverse embedding space to ensure both modality-specific and -shared are effectively used. Building on these assumptions, we formulate our objectives from a mutual information perspective, demonstrating that joint learning of modality-erased and modality-related features optimizes mutual information with identity, producing effective feature representations for mixed-modal matching. Unlike SGEIL [10], which relies on additional shape annotations to remove shape information from identity-aware features and is limited to cross-modal matching, MixER learns both modality-erased and modality-related features in an unsupervised way, enabling it to handle both cross- and mixed-modal tasks.

As illustrated in Fig.1c, MixER leverages modality-erased features for inter-modality matching while refining intra-modality matching by mixing them with modality-related features via ID-modality-aware, modality-confusion, and feature-fusion losses. Furthermore, our framework can be integrated into state-of-the-art approaches to enhance their performance in mixed-modal and cross-modal matching settings.

The main contributions of this paper are summarized as follows. **(1)** We motivate and formalize more comprehensive and challenging settings for mixed-modal V-I ReID evaluation, where galleries may have data from both I and V modalities. **(2)** A mixed modality-erased and -related (MixER) feature learning paradigm is introduced for enhancing mixed- and cross-modal VI-ReID on a single feature embedding backbone. It separates modality-erased features from modality-related ones through orthogonal decomposition and gradient reversal. Modality-erased features capture modality-shared discriminative semantics, while modality-related features are designed to extract additional discriminative attributes unique to each modality, thereby improving the diversity of learned representations. **(3)** Our experiments on SYSU-MM01, RegDB, and LLCM show that MixER strategy outperforms state-of-the-art VI-ReID methods in cross-modal and mixed-modal settings. Additionally, it can be integrated into any VI-ReID method, improving its performance for retrieval applications.

2. Related Work

(a) Cross-modal Person Re-Identification. Person ReID aims to identify individuals by matching distinctive characteristics in query images within a larger gallery [6, 44, 46, 47]. VI-ReID extends this task to low-light conditions, but modality-specific discrepancies pose challenges. Existing methods extract global [1, 20, 31, 33, 38, 40, 49] or part-based features [3, 17, 26, 36, 39] using striping or attention mechanisms. While effective for ID discrimination, these methods struggle to learn modality-invariant representations. Recent approaches focus on modality-invariant feature learning by disentangling identity features from modality-specific attributes [7, 10, 26, 27, 37, 45, 48]. GAN methods [2, 7, 26, 37, 42] attempt to separate content from modality style but often lose identity-related details. [42] generates the intermediate modality to alleviate the modality-specific information. Our MixER separates such information from modality-invariant to make the matching more consistent for inter- and intra-modal scenarios. Other techniques enforce orthogonality between modality-specific and invariant features [11, 45, 48], though ensuring complete modality suppression remains a challenge. Additionally, shape-based approaches [10, 27] use body structure to refine identity features while minimizing modality bias.

(b) Multi-modal Learning. Different data modalities offer complementary features for representation, an area explored in multi-modal learning [24]. Recent transformer-based methods [12, 18, 23] fuse multi-modal inputs directly as tokens rather than extracting separate modality-specific features. Real-world applications, however, often lack complete modality availability, presenting challenges addressed by approaches that handle missing modalities [5, 21, 29, 34]. The ImageBind [14] exemplifies large-scale multi-modal unification, mapping diverse modalities into a shared representation space. In ReID, [22] presents a unified model across RGB, infrared, sketch, and text modalities, achieving strong cross-domain performance. InterReID [30] focuses on multi-modal ReID. Given an image of a person captured in one modality, the objective is to rank similar gallery images in the other modalities. InstructReID [16] further unifies ReID tasks by leveraging a multi-modal backbone for text and image prompts, achieving SOTA performance across various ReID settings. In contrast, we formalize new settings for a comprehensive and realistic mixed-modal ReID evaluation as the general use case for cross-modal person ReID. In particular, it allows us to evaluate the challenge posed by a low modality gap, where the primary difficulty lies in distinguishing individuals of the same modality but different identities.

(c) Critical Analysis. While significant focus has been placed on cross-modal V-I person ReID and its associated challenges, the mixed-modality query-gallery scenario remains largely unaddressed. Despite achieving SOTA results in standard settings, existing VI-ReID methods do not report their performance in mixed-modality settings. [25] proposed a specific mixed-modal setting and proposes a ReID approach that enhances the performance with this setting by applying a re-ranking search on same-modality samples in the gallery. It does not explicitly extract modality-specific information to improve either same- or cross-modal matching. Furthermore, while some methods [32, 48] attempt to disentangle modality-specific features to enhance the modality-shared component in cross-modal contexts, we show that directly using modality-specific information does not improve inter-modal matching, where such information is unavailable and must instead be erased from shared features. These modality-specific features are, however, beneficial for intra-modal matching within mixed-modality settings. Our findings are further supported by mutual information analysis and experimental results across proposed mixed-modal scenarios.

3. The Proposed MixER Method

To address the limitation of ReID methods in mixed-modal settings, we propose the MixER learning paradigm to combine ID-discriminative modality-erased and modality-related features. Fig.2 provides an overview of MixER learning approach. It relies on a shared backbone with three sub-modules to extract independent modality-related and -erased features by applying the orthogonal decomposition, modality-confusion, and modality-related losses.

Problem Definition. A multimodal dataset for VI-ReID is composed of visible $\mathcal{V} = \{x_v^{(j)}, y_v^{(j)}\}_{j=1}^{N_v}$ and infrared $\mathcal{I} = \{x_i^{(j)}, y_i^{(j)}\}_{j=1}^{N_i}$ sets of images from C_y distinct individuals, with their ID labels. Our proposed Mixed VI-ReID system seeks to match images captured from V and I cameras by using one deep backbone model that encodes modality-invariant person embeddings, denoted by $\mathbf{z}_v$ and $\mathbf{z}_i$. Given query images (V or I), the objective is to retrieve images with the same identity over the gallery set containing both V and I modalities, by computing and sorting the distance value $D(.,.)$ for each gallery image:

$$D(\mathbf{z}_m^{(j)}, \mathbf{z}_{m'}^{(p)}) < D(\mathbf{z}_m^{(j)}, \mathbf{z}_{m''}^{(n)}), \tag{1}$$

where $y_m^{(j)} = y_{m'}^{(p)} \neq y_{m''}^{(n)}$, m, m', m'' are modalities that could be v or i independently, and superscripts p and n indicating indices of positive and negative samples, respectively. To learn these features, we decompose them into two independent components, each meeting specific constraints suitable for inter-modal and intra-modal matching. This is achieved by maximizing the mutual information (MI) between these features and the ID labels.

3.1. Mutual Information Analysis

To extract ID-discriminative features from images, the model needs to maximize the MI between these extracted

features[1], Z_m, and label spaces, Y:

$$\max_{Z_m} \quad \mathrm{MI}(Z_m; Y), \tag{2}$$

where Y represents the identity of the individuals in the input images. To ensure that the learned features are effective in mix-modality scenarios, they are decomposed into two independent components: **(a) modality-erased** (Z_m^e), which should not contain any modality information to be proper for inter-modal matching (e.g., short-hair attributes in Fig.1b), and **(b) modality-related** (Z_m^r) component to refine the modality-erased for improving the intra-modal matching. This component should contain ID information that is also relevant to the modality (e.g., text on a t-shirt in Fig.1b). To have two independent components, the MI between them must be zero. Thus, the optimization becomes:

$$\begin{aligned} \max_{Z_m^e, Z_m^r} \quad & \{\mathrm{MI}(Z_m^e, Z_m^r; Y)\} \text{ s.t. } \mathrm{MI}(Z_m^e; Z_m^r) = 0, \\ & \mathrm{MI}(Z_m^e; M) = 0 \text{ and } \mathrm{MI}(Z_m^r; Y|M) = 0, \end{aligned} \tag{3}$$

where M is the modality label space. Constraint $\mathrm{MI}(Z_m^e; M)$=0 ensures that modality information is erased from Z_m^e and $\mathrm{MI}(Z_m^r; Y|M)$=0 ensures that Z_m^r does not contain ID-aware information, which disregards the modality label. Also, Z_m^e and Z_m^r should be independent.

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

Proof. Can be found in the supplementary materials. □

By incorporating M into Eq. (3) and using Theorem 1:

$$\begin{aligned} \mathrm{MI}(Z_m^e, Z_m^r; Y) =& \mathrm{MI}(Z_m^e; Y) + \mathrm{MI}(Z_m^r; Y) \\ =& \mathrm{MI}(Z_m^e; Y|M) + \mathrm{MI}(Z_m^e; Y; M) \\ &+ \mathrm{MI}(Z_m^r; Y|M) + \mathrm{MI}(Z_m^r; Y; M). \end{aligned} \tag{4}$$

Since the MI(.;.) is non-negative, $\mathrm{MI}(Z_m^e; Y; M)$=0, and $\mathrm{MI}(Z_m^r; Y|M)$=0, Eq. (3) can be formulated as the maximization of the following Lagrangian:

$$\max_{Z_m^e, Z_m^r} \{\overbrace{\mathrm{MI}(Z_m^e; Y|M) - \lambda_1 \mathrm{MI}(Z_m^e; M)}^{\textit{Modality-Erased Learning}} + \underbrace{\mathrm{MI}(Z_m^r; Y; M)}_{\textit{Modality-Related Learning}} - \underbrace{\lambda_2 \mathrm{MI}(Z_m^e; Z_m^r)}_{\text{Orthogonal Feature Learning}} \}. \tag{5}$$

3.2. Modality Erased and Related Learning

(a) Orthogonal Feature Decomposition. To decompose the extracted features into modality-erased and -related feature vectors, two modality-specific and one shared sub-modules are proposed to project the $\mathbf{z}_m$ to them as:

$$\mathbf{z}_m^e = \mathcal{F}_s(\mathbf{z}_m), \quad \mathbf{z}_m^r = \mathcal{F}_m(\mathbf{z}_m), \tag{6}$$

[1]Uppercase is used as random variables and lowercase for samples.

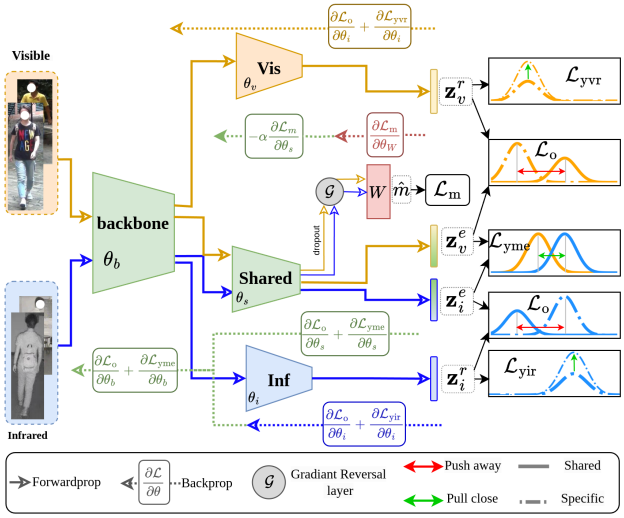


Figure 2. The overall architecture of our proposed MixER method. It extracts two independent ID-discriminative feature vectors by orthogonal decomposition, modality-confusion, and modality-aware losses for learning modality-erased and modality-related feature representation.

where $\mathbf{z}_m = \mathcal{F}_b(x_m)$ is extracted by a shared backbone. One reason behind using specific sub-modules for modality-related features is to prevent them from sharing information through model parameters.

Minimizing $\mathrm{MI}(Z_m^e; Z_m^r)$: When minimizing the MI between Z_m^r and Z_m^e, they must be independent to avoid affecting each other through the learning. Since the MI estimation is complex and time-consuming [4, 10], we estimate the independence as minimizing the orthogonal loss:

$$\mathcal{L}_{\mathrm{o}} = \mathbb{E}_{(\mathbf{z}_m^r, \mathbf{z}_m^e) \sim (Z_m^r, Z_m^e)} \frac{\mathbf{z}_m^r \cdot \mathbf{z}_m^e}{\|\mathbf{z}_m^r\|_2 \, \|\mathbf{z}_m^e\|_2}. \tag{7}$$

(b) Learning Modality Erased Features. The goal of modality-erased learning is extracting features, $\mathbf{z}_m^r$, that discriminate the ID without using modality information to make them appropriate for cross-modal matching when the modality-specific information is absent.

Maximizing $\mathrm{MI}(Z_m^e; Y|M)$: given the MI attributes, we expand it as (The proof is in the suppl. materials):

$$\mathrm{MI}(Z_m^e; Y|M) = \mathrm{MI}(Z_m^e; Y) - \mathrm{MI}(Z_m^e; Y; M) = \mathrm{MI}(Z_m^e; Y). \tag{8}$$

To maximize Eq. (8), the cross-entropy loss $\mathcal{L}_{\mathrm{ce}}$ must be minimized as (see the property **??** in supply. materials):

$$\mathcal{L}_{\mathrm{meid}}(Z_m^e, Y) = \mathbb{E}_{(\mathbf{z}_m^e, y_m) \sim (Z_m^e, Y)} \mathcal{L}_{\mathrm{ce}}(\mathbf{z}_m^e, y_m). \tag{9}$$

To obtain distinct features for each person, the center-cluster loss($\mathcal{L}_{\mathrm{cc}}$) [36] is also minimized. So, our modality-erased identity loss is:

$$\mathcal{L}_{\mathrm{yme}} = \mathcal{L}_{\mathrm{meid}}(Z_m^e, Y) + \mathcal{L}_{\mathrm{cc}}(Z_m^e, Y). \tag{10}$$

Minimizing MI($Z_m^e; M$): Modality-erased features should not distinguish the origin modality of input samples. Therefore, an adversarial modality classifier based on the Gradient Reversed Layer (GRL)[13] is used to propagate the reverse gradient onto model parameters. The modality-confusion loss for this objective is:

$$\mathcal{L}_{\text{m}} = \mathbb{E}_{(\mathbf{z}_m^e, m)\sim(Z_m^e, M)} \mathcal{L}_{\text{ce}}(\mathcal{G}(W^T \mathbf{z}_m^e), m), \tag{11}$$

where $\mathcal{G}$ is the GRL and $W \in \mathbb{R}^{d\times 2}$ is a linear layer.

(c) Learning Modality Related Features. To learn features that leverage modality-related ID-discriminating information simultaneously, a new doubled label space is proposed to account for identity alongside the modality by separating the same person in each modality:

$$y'_m \sim Y' = \begin{cases} 2y_m & m = \text{v} \\ 2y_m + 1 & m = \text{i}, \end{cases} \tag{12}$$

and minimizing the cross-entropy loss between $\mathbf{z}_m^r$ and y':

$$\mathcal{L}_{\text{mrid}}(Z_m^r, Y') = \mathbb{E}_{(\mathbf{z}_m^r, y'_m)\sim(Z_m^r, Y')} \mathcal{L}_{\text{ce}}(\mathbf{z}_m^r, y'_m). \tag{13}$$

The modality-aware loss function for Z_m^r is :

$$\mathcal{L}_{\text{ymr}} = \mathcal{L}_{\text{mrid}}(Z_m^r, Y') + \mathcal{L}_{\text{cc}}(Z_m^r, Y'), \tag{14}$$

where $\mathcal{L}_{\text{cc}}$ is the center-cluster loss [36].

(d) Mixed-Modal Matching Fusion. To balance modality-erased and modality-related features for the matching process at the inference time, we propose a mixed cross-modal triplet loss to avoid having one component be dominant or useless. The feature fusion loss is:

$$\begin{aligned} \mathcal{L}_{\text{f}} = &\max\{D(\mathbf{z}_m^{f,(j)}, \mathbf{z}_m^{f,(p)}) - D(\mathbf{z}_m^{e,(j)}, \mathbf{z}_{\tilde{m}}^{e,(n)}) + \alpha, 0\} \\ &+ \max\{D(\mathbf{z}_m^{e,(j)}, \mathbf{z}_{\tilde{m}}^{e,(p)}) - D(\mathbf{z}_m^{f,(j)}, \mathbf{z}_m^{f,(n)}) + \alpha, 0\}, \end{aligned} \tag{15}$$

where $D(.,.)$ is the distance between two embeddings, $\tilde{m} \neq m$, n and p are positive and negative samples to the j instance. $\mathbf{z}_m^f$ is formed by concatenating $\mathbf{z}_m^e$ and $\mathbf{z}_m^r$.

(e) Overall Training. We jointly optimize the network in an end-to-end manner by using the overall loss:

$$\mathcal{L} = \mathcal{L}_{\text{yme}} + \mathcal{L}_{\text{ymr}} + \lambda_{\text{m}}\mathcal{L}_{\text{m}} + \lambda_{\text{o}}\mathcal{L}_{\text{o}} + \lambda_{\text{f}}\mathcal{L}_{\text{f}}, \tag{16}$$

where λ_{m}, λ_{o} and λ_{f} are loss-weighting hyperparameters.

(f) Inference. During inference, MixER performs mixed-modal matching using a single backbone with three small projection heads rather than using three different backbones. For a given image x_m, it extracts $\mathbf{z}_m^r$, $\mathbf{z}_m^e$ and then $\mathbf{z}_f^r$. Matching scores are computed using cosine similarity between feature vectors. Fused features ($\mathbf{z}_f^r$), are used for images within the same modality, while cross-modal matching uses only modality-erased features ($\mathbf{z}_m^e$).

4. Results and Discussion

4.1. Experimental Methodology

(a) Datasets. Research on cross-modal V-I ReID has mainly used the SYSU-MM01 [35], RegDB [28] datasets, and LLCM [41] datasets. SYSU-MM01 is a large dataset containing more than 22K V and 11K I images of 491 individuals captured with 4 RGB and 2 NIR cameras. RegDB contains 4K co-located V-I images of 412 individuals. Randomly divide the dataset into two sets of the same size for training and testing. The LLCM dataset consists of a low-light cross-modality dataset comprising 1K identities and is divided into training and testing sets at a 2:1 ratio.

(b) Mixed-Modal Evaluations. Since mixed-modal evaluation has not been performed on V-I datasets, we introduce multiple settings of such assessment based on real-world scenarios where the query and gallery images can be either I or V. Unlike cross-modality settings, where the query and gallery sets strictly contain images of opposite modalities, the mixed-modal setting may contain images from both modalities in the query and gallery sets. We structured the mixed-modality settings into 5 cases as illustrated in Fig.3, each one defined by the exclusion of images from the gallery based on the query identity and camera: **(1) Mix**: Images of all individuals and cameras are included in the gallery, regardless of the query camera type. **(2) Mix-Camera**: The gallery excludes images from the query's camera and identity. **(3) Mix-Camera-ID**: The gallery excludes images taken by the same camera as the query and belonging to the query person. **(4) Mix-ID**: Excludes all images of the query person captured by cameras of the same modality as the query. **(5) Mix-P**: Includes the images of the same person ID, seen in the same modality in the query and gallery sets [25]. We argue that this is a relatively easy and less realistic setting compared to the first 4 in the list, which diminishes its relevance as a research objective.

Since ReID methods have not been evaluated in these settings, we retrained several open-sourced state-of-the-art (SOTA) VI-ReID methods to assess their performance across these configurations. Rank-1 (R1) accuracy and mean average precision (mAP) were measured for each setting in line with the dataset evaluation criteria. Note that for the RegDB and LLCM datasets, there is only one camera per modality. Therefore, the "Mix-Camera" and "Mix-Camera-Identity" settings are not applicable.

(c) Implementation Details. To extract modality-erased features, the SAAI model[9] was used without prototype learning as a baseline with ResNet50 [15] as the backbone. For modality-related learning modules, `layer4` is cloned from the backbone for each modality. Each image input is resized to 288 by 144, then cropped and erased randomly, and filled with zero padding or mean pixels. We used an ADAM optimizer [19] with a linear warm-up strategy for

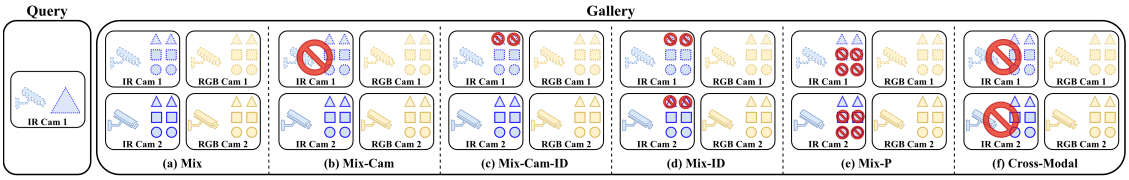


Figure 3. Different settings for forming a gallery based on modality, camera, and query identity. The gallery set images are: (a) all images from both modalities; (b) and all images except ones captured with the same camera; (c) same camera and same identity; (d) same identity in the same modality as the query; (e) is mixed settings introduced in [25] and (f) all images from another modality.

the optimization process. Each training batch contains 8 V and 8 I images from 10 randomly selected identities. The model was trained for 180 epochs, following [9], the initial learning rate is set to 0.0004 and decreased by factors of 0.1 and 0.01 at 80 and 120 epochs, respectively.

(d) Benchmark methods. To assess effectiveness across different settings, we benchmarked MixER against several open-source SOTA VI-ReID methods—DDAG [39], MPANet [36], DEEN [41], SGEIL [10], SAAI [9], and IDKL [32]. Each method was retrained from scratch using the authors' provided parameters, with implementation details available in the supplementary materials.

4.2. Comparison with State-of-the-Art Methods

(a) Mixed-Modal Results. Tables 1, 2a, and 2b show the performance of SOTA alongside our MixER across various mixed-modal settings with an I query (V query results are in suppl. materials) on SYSU-MM01(single-shot), RegDB, and LLCM, respectively. MixER(E) and MixER(R) use only $\mathbf{z}_m^e$ and $\mathbf{z}_m^r$ features, respectively, while MixER in a cross-modal setting uses $\mathbf{z}_m^e$ and $\mathbf{z}_m^f$ for the mixed-modal cases. These cross-dataset results highlight the effectiveness of our approach across different VI-ReID methods. In these mixed settings, the difficulty level increases from setting 1 to 4 due to the progressive reduction in positive samples within the same modality as the query, challenging each method's robustness in mixed scenarios. Notably, this analysis also enables us to examine each method's strengths and weaknesses in novel contexts; for instance, SGEIL [10] performs best in the "MIX" but exhibits a significant performance drop in the "Mix-ID" on the SYSU-MM01 dataset compared to other methods. In contrast to Mix-P [25], where the R1 lacks discriminative power for comparing different VI-ReID approaches due to their uniformly high performance, our experimental settings offer a more comprehensive and realistic framework for evaluating VI-ReID methods. Since the source code of [25] is unavailable, we could not directly evaluate it under our settings. The proposed MixER consistently outperforms existing methods across nearly all mixed settings, highlighting its capacity to bridge the gap between V and I images by learning modality-erased features while enhancing same-modality matching through modality-related information. Importantly, modality-related features (MixER(R)) perform less effectively in Mix-ID and Cross-modal settings, where the same person's modality does not appear in the gallery. However, they refine intra-modal matches by adjusting similarity scores, supporting the exclusion of modality-specific information for robust inter-modal alignment.

To further evaluate the adaptability of our learning paradigm, we applied it to several SOTA VI-ReID methods as a baseline on the SYSU dataset. We tested it in two variations: (a) stopping gradient backpropagation of our proposed modality-related and erased learning losses and (b) enabling end-to-end training with combined gradients from the existing and our proposed losses to strengthen cross-modal matching by filtering out modality-related features from the backbone. Results in Table 3 indicate

Method	Venue	Mixed-Modal										Cross-Modal				Uni-Modal			
		Mix		**Mix-Cam**		**Mix-Cam-ID**		**Mix-ID**		Mix-P[25]		**All**		**Indoor**		**I→I**		**V→V**	
		R1	**mAP**	**R1**	**mAP**	**R1**	**mAP**	**R1**	**mAP**	**R1**	**mAP**	**R1**	**mAP**	**R1**	**mAP**	**R1**	**mAP**	**R1**	**mAP**
DDAG[39]	ECCV20	91.91	57.60	80.81	53.54	70.41	45.14	23.92	29.09	98.41	45.87	53.29	50.61	61.02	67.98	80.15	86.80	97.89	91.90
MPANet [36]	CVPR21	95.46	72.82	89.45	70.34	81.95	63.81	49.06	51.08	99.16	72.87	70.58	68.24	76.54	80.95	88.53	92.85	98.31	94.08
DEEN[41] + Flip	CVPR23	-	-	-	-	-	-	-	-	-	-	74.7	71.8	80.3	83.3	-	-	-	-
DEEN[41]	CVPR23	94.07	74.06	87.50	72.09	79.50	65.66	56.77	55.30	98.51	74.06	72.55	68.59	84.21	82.14	86.14	90.98	96.83	89.93
SGEIL[10]	CVPR23	94.82	71.03	89.22	70.16	79.33	61.34	46.60	48.37	99.76	70.13	73.34	67.44	84.09	81.65	86.52	91.36	97.29	90.24
SAAI[9] +AIM	ICCV23	-	-	-	-	-	-	-	-	-	-	75.90	77.03	83.20	88.01	-	-	-	-
SAAI[9]	ICCV23	96.01	74.59	90.63	72.51	84.30	65.94	52.49	53.30	99.17	72.55	73.87	69.71	84.19	82.59	89.29	93.06	98.24	93.46
IDKL[32]	CVPR24	95.27	72.81	90.15	71.87	80.73	63.54	47.34	50.86	99.53	72.79	72.05	69.67	84.54	83.76	89.55	93.18	98.55	95.13
CIDHL [25]	-	-	-	-	-	-	-	-	-	90.81	53.24	-	-	-	-	-	-	-	-
MixER(R)	Ours	96.22	38.94	88.32	25.37	82.58	23.65	0.87	0.99	99.99	38.94	1.19	3.61	1.94	5.2	90.02	93.93	98.63	94.81
MixER(E)		96.46	75.21	90.24	72.89	84.55	66.02	52.89	54.86	98.24	76.22	75.04	71.22	85.82	84.06	90.18	93.65	97.95	93.38
MixER		**96.63**	**79.64**	**91.77**	**76.35**	**87.56**	**72.70**	**65.14**	**62.76**	98.89	80.39	75.04	71.22	85.82	84.06	**93.06**	**95.82**	**99.14**	**95.30**
MixER +AIM		-	-	-	-	-	-	-	-	-	-	76.12	75.45	85.30	85.54	-	-	-	-

Table 1. Performance of SOTA VI-ReID techniques in mixed, cross, and uni-modal settings on the SYSU-MM01 dataset. AIM (Affinity Inference) is a re-ranking process originally applied in the SAAI [9] GitHub projects. Notably, the Mix-P setting [25] yields highly accurate results, suggesting its lower difficulty and limited relevance as a realistic evaluation.

Method	Mixed-Modal				Cross-Modal			
	Mix		Mix-ID		I→V		V→I	
	R1	mAP	R1	mAP	R1	mAP	R1	mAP
DDAG [39]	99.9	76.17	45.29	45.54	69.34	63.46	68.06	61.80
MPANet [36]	**100**	76.46	42.18	44.66	84.27	80.20	83.20	79.82
DEEN [41]	99.95	83.06	66.21	59.61	90.29	83.98	91.21	85.13
SAAI [9]	**100**	82.29	58.01	55.41	86.21	80.0	86.60	81.51
IDKL [32]	**100**	83.83	61.46	60.54	87.23	83.20	87.91	85.07
SAAI +AIM	-	-	-	-	92.09	92.01	91.07	91.45
IDKL + KR	-	-	-	-	94.22	90.43	94.72	90.19
MixER	**99.9**	**89.44**	**79.95**	**73.45**	90.49	85.63	90.53	86.42
MixER + AIM	-	-	-	-	90.78	90.18	90.35	90.02
MixER + KR	-	-	-	-	**97.09**	**93.01**	**96.55**	**93.55**

(a) RegDB dataset.

Method	Mixed-Modal				Cross-Modal			
	Mix		Mix-ID		I→V		V→I	
	R1	mAP	R1	mAP	R1	mAP	R1	mAP
DDAG [39]	69.34	63.46	68.06	61.80	40.14	26.88	45.17	29.94
MPANet [36]	96.76	53.73	38.79	29.03	46.48	30.96	52.15	37.00
DEEN [41]	97.63	**64.69**	49.11	38.38	69.49	54.75	73.95	58.75
SAAI [9]	96.61	60.05	40.86	31.27	59.37	45.65	64.37	48.60
IDKL [32]	97.73	62.46	38.88	33.85	62.53	49.33	70.36	55.04
IDKL + KR	-	-	-	-	70.72	65.19	72.22	66.43
MixER	**97.80**	64.45	**57.1**	**45.71**	65.76	51.08	70.79	56.61
MixER+ AIM	-	-	-	-	66.11	62.89	74.10	**68.20**
MixER + KR	-	-	-	-	**73.72**	**65.36**	**76.14**	65.46

(b) LLCM dataset.

Table 2. Performance of SOTA VI-ReID techniques in mixed, cross, and uni-modal settings on the (a) RegDB and (b) LLCM datasets.

that modality-related learning significantly boosts performance in mixed-modal settings. In contrast, end-to-end modality-erased learning enhances both mixed and cross-modal matching. For example, our modality-related learning improves the mAP of SAAI[9] and IDKL [32] in "Mix-ID" settings by more than 6% and 4%, respectively.

(b) Cross-Modal Results. To show that erasing modality-specific information from features makes them more proper for cross-modal matching, we measure the performance of MixER and compare it with SOTA VI-ReID in Tables 1, 2a, and 2b. Our experiments show that MixER outperforms these methods in various situations. Modality-specific sub-modules are not triggered during inference in cross-modal matching. For example, compared to the second-best approach for the "All Search" scenario, MixER outperforms by a margin of 2.3% R1 and 3.3% mAP without adding complexity to the model. For the "I→V" mode on RegDB, MixER achieves 91.4% R1 accuracy and 85.5% mAP. For the "V→I" mode, our method also obtains 90.4% R1 accuracy and 86.8% mAP. The results validate the effectiveness of our proposed MixER and show that it can effectively reduce the discrepancy between the V and I modalities.

In addition, our modality-erased feature learning has the advantage that it can be used in different VI-ReID models to improve their performance in cross-modal settings without incurring overhead during testing. To show this adaptability, in Table 3, in the first column, end-to-end training improves the performance of all methods in the SYSU-MM01 dataset. For example, MixER increases the mAP of SGEIL[10] and SAAI[9] by 1.75% and 1.37%.

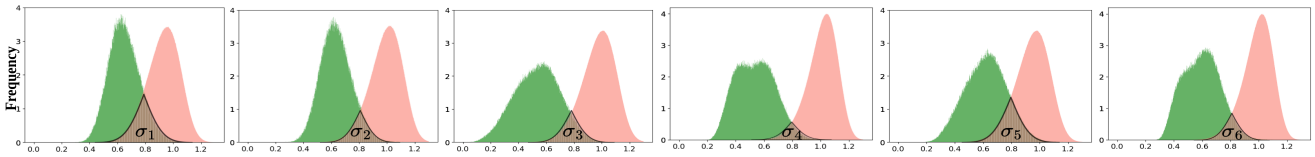
(c) Uni-Modal Results. To illustrate the effectiveness of learning modality-related alongside modality-erased in MixER for uni-modal matching, our trained model was tested on the SYSU-MM01 dataset in V→V and I→I settings, as well as on an unseen V dataset. Table 1 (column "Uni-Modal") presents these results for SYSU-MM01, demonstrating that our MixER approach outperforms other methods since it uses the complementary information of discriminative modality-related and modality-erased features. Table 4 reports the performance of various methods, while each column indicates the V-I training dataset, and each row shows the model's performance on the Mar-

Method	Cross-modal		Mix		Mix-Cam		Mix-Cam-ID		Mix-ID	
	R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP
DDAG[39]	53.29	50.61	91.91	57.60	80.81	53.54	70.41	45.14	23.92	29.09
DDAG(f)+MixER	53.21	50.86	92.04	60.29	81.27	55.07	72.05	48.68	35.71	34.81
DDAG+MixER	54.61	51.27	92.52	62.52	81.74	57.16	72.96	51.93	38.79	38.79
MPANet[36]	69.34	66.42	95.46	72.82	89.45	70.34	81.95	63.81	49.06	51.08
MPANet(f)+MixER	69.66	66.51	95.02	77.01	90.43	72.63	83.98	69.80	63.96	60.75
MPANet+MixER	72.18	68.18	95.80	78.67	91.34	74.98	83.95	70.55	66.00	61.16
DEEN[41]	72.55	68.59	94.07	74.06	87.50	72.09	79.50	65.66	56.77	55.30
DEEN(f)+MixER	72.63	68.49	94.55	78.91	88.38	75.27	82.77	70.9	65.8	60.05
DEEN+MixER	74.19	69.00	95.00	79.00	89.00	76.00	83.00	71.00	66.00	61.00
SGEIL[10]	73.34	67.44	94.82	71.03	89.22	70.16	79.33	61.34	46.60	48.37
SGEIL(f)+MixER	72.17	67.81	94.90	76.45	90.14	73.43	82.63	68.86	64.14	60.15
SGEIL+MixER	74.08	69.19	95.33	78.01	91.16	74.29	83.97	69.40	65.37	61.50
SAAI[9]	72.19	69.71	96.01	74.59	90.63	72.51	84.30	65.94	52.49	53.30
SAAI(f)+MixER	73.64	69.51	96.21	79.12	91.7	75.03	87.0	71.22	63.72	60.91
SAAI+MixER	75.37	71.92	97.27	80.47	92.66	76.81	88.51	73.24	66.23	62.45
IDKL[32]	72.05	69.67	95.27	72.81	90.15	71.87	80.73	63.54	47.34	50.86
IDKL(f)+MixER	72.65	70.2	95.58	74.24	90.51	71.78	83.77	65.41	52.63	55.57
IDKL+MixER	73.44	71.02	96.10	76.67	91.38	74.27	85.61	68.44	58.37	57.91

Table 3. Accuracy of the proposed MixER and SOTA methods as baseline on SYSU-MM01 (single-shot setting) as I query. "(f)" indicates that MixER losses were not back-propagated through baseline models.

Source:	SYSU-MM01		RegDB		LLCM	
Target:	Market1501(V→V)					
	R1	mAP	R1	mAP	R1	mAP
DDAG[39]	82.39	55.57	11.99	3.32	53.40	20.18
MPANet[36]	78.97	51.61	**18.17**	4.79	56.91	22.79
DEEN [41]	66.50	33.81	6.85	1.7	58.81	23.72
SGEIL [10]	79.18	48.72	-	-	-	-
SAAI [9]	84.32	57.62	14.54	3.44	55.68	22.77
IDKL[32]	76.66	49.59	12.84	2.79	54.61	21.95
MixER(E)	82.90	56.56	14.31	3.97	54.45	22.52
MixER(R)	83.46	56.85	16.29	4.08	59.56	25.93
MixER(E+R)	**87.93**	**62.41**	17.7	**5.02**	**61.37**	**27.88**
Upper-bound	95.1	87.8	95.1	87.8	95.1	87.8

Table 4. Performance of V-I ReID methods on Cross-Dataset RGB Market1501 dataset. Columns indicate the training dataset, and rows show model performance on Market1501.



(a) Baseline (Cross) (b) MixER (Cross) (c) Baseline (Mix) (d) MixER (Mix) (e) Baseline (Mix-Cam) (f) MixER (Mix-Cam)

Figure 4. The intra-class (green) and inter-class (pink) distances distribution of features in different gallery settings. SAAI is the baseline.

ket1501 [43] dataset. Our MixER approach achieves the highest R1 and mAP scores, with values of 87.9% and 62.4%, respectively. When using only modality-related features (activating only the V branch) or only modality-erased features, the mAP scores are 56.8% and 56.5%, respectively. However, the integration of both feature types increases the mAP and R1 scores by more than 3%, indicating better generalization and complementary representations.

4.3. Ablation Studies

Losses. Table 5 shows the impact of each loss component on performance. Using only the $\mathcal{L}_{\text{y}}$ as a baseline achieves results of 69.7% R1 and 66.3% mAP in cross-modal matching. Adding $\mathcal{L}_{\text{ymr}}$ improves Mix-Cam-ID R1 to 83.6%, indicating $\mathcal{L}_{\text{ymr}}$'s benefit for intra-modality matching consistency, though cross-modal performance slightly decreases. Incorporating only $\mathcal{L}_{\text{o}}$ improves both cross- and mixed-modal performance, whereas $\mathcal{L}_{\text{m}}$ provides a more significant enhancement in Cross-Modal metrics. The combined effect of $\mathcal{L}_{\text{y}}$, $\mathcal{L}_{\text{yme}}$, $\mathcal{L}_{\text{o}}$, and $\mathcal{L}_{\text{m}}$ achieves 73.4% R1 in cross-modal, showing these losses' synergy. Finally, adding $\mathcal{L}_{\text{f}}$ achieves optimal results across all metrics, highlighting its important role in refining features for mixed-modal settings.

Settings					Cross-Modal		Mix-Cam-ID	
$\mathcal{L}_{\text{yme}}$	$\mathcal{L}_{\text{ymr}}$	$\mathcal{L}_{\text{o}}$	$\mathcal{L}_{\text{m}}$	$\mathcal{L}_{\text{f}}$	**R1**	**mAP**	**R1**	**mAP**
✓					69.74	66.38	80.57	62.51
✓	✓				69.28	66.09	83.61	64.57
✓		✓			70.25	67.11	81.07	62.99
✓			✓		71.72	67.60	78.55	60.38
✓		✓	✓		72.30	68.05	82.3	63.76
✓	✓	✓	✓		73.40	70.87	84.64	65.83
✓	✓	✓	✓	✓	**73.43**	**70.92**	**87.56**	**72.70**

Table 5. Impact of losses on MixER performance.

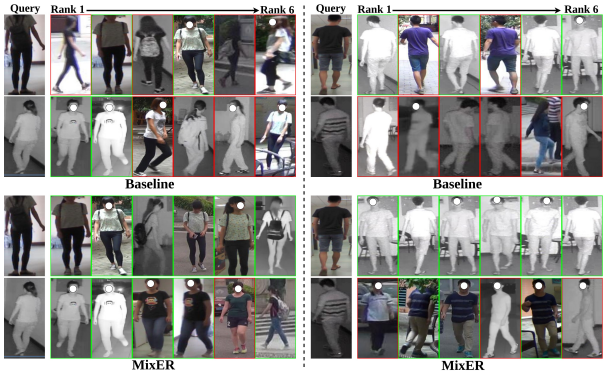


Figure 5. Rank-6 retrieval for baseline (top) and MixER (bottom) on SYSU in Mix-Cam-ID (left) and Mix-ID (right) settings.

4.4. Qualitative Analysis

Feature distribution. To examine the effectiveness of our method in cross-modal and mixed-modal matching, we visualize inter-class and intra-class distance distributions on the SYSU dataset, as shown in Fig.4(a-f). Compared to the baseline, MixER reduces both the mean and variance of intra-class distances across all settings, decreasing the area of overlap ($\sigma_2 < \sigma_1$, $\sigma_4 < \sigma_3$, and $\sigma_6 < \sigma_5$) showing that it more effectively reduces intra-class distances, thereby decreasing modality discrepancy in modality-erased features and enhancing identity discrimination in modality-related features. Also, UMAP visualizations (see Appendix) show that MixER better separates identities and reduces modality discrepancy within the mixed-modal gallery.

Retrieval results. To show the effectiveness of MixER, we present retrieval results on the SYSU-MM01 dataset in Fig. 5. Green boxes indicate correct matches, while red boxes represent incorrect ones. Overall, combining modality-erased and -related features at the bottom significantly improves the results, with more correct matches appearing in higher ranked positions compared to the baseline. Modality-erased learning aims to find the closest matches based solely on shared attributes between V and I images. However, it may mistakenly select incorrect images. Modality-related features refine this by penalizing incorrect matches based on attributes specific to the query modality. For example, in the top second row, the model retrieves images of a person in a T-shirt and shorts. In the bottom row, it produces better matches by using modality-erased and modality-related features.

5. Conclusion

In this paper, we propose Mix-Modal ReID, a new evaluation setting that better reflects real-world person re-identification challenges in mixed visible-infrared galleries. To enhance robustness across modalities, we introduce a modality-erased and modality-related feature learning approach. Experiments on SYSU-MM01, RegDB, and LLCM show that our method can outperform state-of-the-art VI-ReID approaches in both mixed- and cross-modal settings, revealing the limitations of existing methods. Our approach provides a strong foundation for more effective and adaptable VI-ReID systems in real-world applications.

Acknowledgment: This research was supported by the Natural Sciences and Engineering Research Council of Canada, and the Digital Research Alliance of Canada.

References

[1] 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

[2] 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

[3] 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

[4] 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

[5] 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

[6] 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

[7] 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

[8] 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

[9] 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

[10] 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

[11] 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

[12] 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

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

[14] 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

[15] 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

[16] 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

[17] 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

[18] 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

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

[20] 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

[21] 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

[22] 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

[23] 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

[24] 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

[25] 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

[26] 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

[27] 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

[28] 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

[29] 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

[30] 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

[31] 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

[32] 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

[33] 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

[34] 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

[35] 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

[36] 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

[37] 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

[38] 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

[39] 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

[40] 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

[41] 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

[42] 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

[43] 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

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

[45] 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

[46] 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

[47] 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

[48] 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

[49] 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