# Learning by Aligning: Visible-Infrared Person Re-identification using Cross-Modal Correspondences

Hyunjong Park* Sanghoon Lee* Junghyup Lee Bumsub Ham†
School of Electrical and Electronic Engineering, Yonsei University
https://cvlab.yonsei.ac.kr/projects/LbA

## Abstract

*We address the problem of visible-infrared person re-identification (VI-reID), that is, retrieving a set of person images, captured by visible or infrared cameras, in a cross-modal setting. Two main challenges in VI-reID are intra-class variations across person images, and cross-modal discrepancies between visible and infrared images. Assuming that the person images are roughly aligned, previous approaches attempt to learn coarse image- or rigid part-level person representations that are discriminative and generalizable across different modalities. However, the person images, typically cropped by off-the-shelf object detectors, are not necessarily well-aligned, which distract discriminative person representation learning. In this paper, we introduce a novel feature learning framework that addresses these problems in a unified way. To this end, we propose to exploit dense correspondences between cross-modal person images. This allows to address the cross-modal discrepancies in a pixel-level, suppressing modality-related features from person representations more effectively. This also encourages pixel-wise associations between cross-modal local features, further facilitating discriminative feature learning for VI-reID. Extensive experiments and analyses on standard VI-reID benchmarks demonstrate the effectiveness of our approach, which significantly outperforms the state of the art.*

## 1. Introduction

Person re-identification (reID) aims at retrieving person images, captured across multiple cameras, with the same identity as a query person. It provides a wide range of applications, including surveillance, security, and pedestrian analysis, and has gained a lot of attention over the last decade [40, 47]. Most reID methods formulate the task as a single-modality retrieval problem, and focus on finding matches, *e.g.*, between RGB images. Visible cameras are incapable of capturing appearances of persons, particularly important for person reID, under poor-illumination conditions (*e.g.*, at night time or dark indoors). Infrared (IR) cameras, on the other hand, work well, regardless of visual light, capturing an overall scene layout, while not taking scene details, such as texture and color. Accordingly, visible-IR person re-identification (VI-reID), that is, retrieving IR person images of the same identity as an RGB query and vice versa, has recently been of great interest [36].

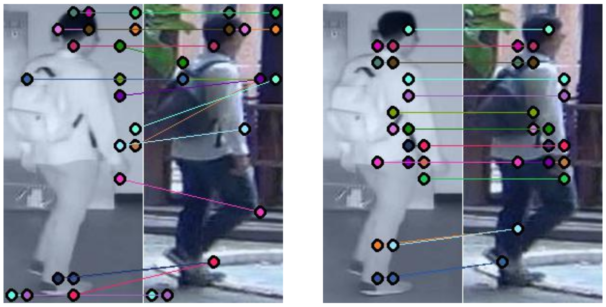


Figure 1: An example of dense cross-modal correspondences between RGB and IR images from the SYSU-MM01 dataset [36]. For the visualization purpose only, we show the top 20 matches according to similarities between local person representations learned without (left) and with (right) our approach. Our person representations are robust to the cross-modal discrepancies, while being highly discriminative, especially for person regions. (Best viewed in color.)

VI-reID is extremely challenging due to intra-class variations (*e.g.* viewpoint, pose, illumination and background clutter), noisy samples (*e.g.* misalignment and occlusion), and cross-modal discrepancies between RGB and IR images. Visual attributes and statistics of RGB/IR images are significantly different from one another [36]. VI-reID methods based on convolutional neural networks (CNNs) alleviate the discrepancies using cross-modal metric losses [7, 37, 41] along with a modality discriminator [4] to learn person representations robust to the cross-modal discrepancies, and further refine the representations with self-attention [39] or disentanglement techniques [3]. These approaches focus on learning coarse image-level or rigid part-level representations, assuming that person images are roughly aligned.

[*] Equal contribution. [†] Corresponding author.

Misaligned features from RGB and IR images, however, have an adverse effect on handling the cross-modal discrepancies, distracting learning person representations.

In this paper, we propose to leverage dense correspondences between cross-modal images during training for VI-reID. To this end, we encourage person representations of RGB images to reconstruct those from IR images of the same identity, which often depict different appearances due to viewpoint and pose variations, and vice versa. We achieve this by establishing dense cross-modal correspondences between RGB and IR person images in a probabilistic way. We incorporate parameter-free person masks to focus on the reconstructions of person regions, while discarding others including background or occluded regions. We also introduce novel ID consistency and dense triplet losses using pixel-level associations, allowing the network to learn more discriminative person representations. Dense cross-modal correspondences align pixel-level person representations from RGB and IR image explicitly, which is beneficial to person representation learning for VI-reID due to two main reasons. First, by enforcing semantically similar regions from RGB and IR images to be embedded nearby, we encourage the network to extract features invariant to the input modalities, even from misaligned RGB and IR person images. Second, by encouraging a local association, we enforce the network to focus on extracting discriminative pixel-wise local features, which further facilitates the person representation learning. The network trained using our framework is thus able to offer local features that are robust to cross-modal discrepancies and highly discriminative (Fig. 1), which are aggregated to form a final person representation for VI-reID, without any additional parameters at test time. Experimental results and extensive analyses on standard VI-reID benchmarks demonstrate the effectiveness and efficiency of our approach. The main contributions of this paper can be summarized as follows:

- We propose a novel feature learning framework for VI-reID using dense cross-modal correspondences that alleviates the discrepancies between multi-modal images effectively, while further enhancing the discriminative power of person representations.
- We introduce ID consistency and dense triplet losses to train our network end-to-end, which help to extract discriminative person representations using cross-modal correspondences.
- We achieve a new state of the art on standard VI-reID benchmarks and demonstrate the effectiveness and efficiency of our approach through extensive experiments with ablation studies.

## 2. Related work

In this section, we briefly describe representative works related to ours, including person reID, VI-reID, cross-modal image retrieval and dense correspondence.

**ReID.** Person reID methods typically tackle a single-modality case, that is, RGB-to-RGB matching. They formulate the reID task as a multi-class classification problem [49], where person images of the same identity belong to the same category. A triplet loss is further exploited to encourage person representations obtained from the same identity to be embedded nearby, while those from the different identities to be distant in feature space [12]. Recent methods focus on extracting person representations robust to intra-class variations, exploiting attributes to offer complementary information [19], disentangling identity-related features [9, 48], or incorporating attention techniques to see discriminative regions [18, 46]. Many reID methods leverage part-based representations [8, 32, 33], which further enhance the discriminative power of person features. Specifically, they divide person images into multiple horizontal grids exploiting human body parts implicitly. Local features from the horizontal parts are more robust to intra-class variations, especially for occlusion, than the global one. However, when body parts from corresponding horizontal grids are misaligned, this rather distracts learning person representations. The works of [15, 23, 45, 52] propose to align semantically related regions between person images, by employing auxiliary pose estimators [23, 45] or human semantic parsing techniques [15, 52]. While these auxiliary branches offer reliable estimations to guide the alignment, they have two main drawbacks: First, they typically require additional datasets during training. Second, the auxiliary predictions are required at test time, making the overall pipeline computationally heavy. On the other hand, we perform the alignment during training only, by leveraging dense correspondences, without additional supervisory signals except ID labels, while enabling an efficient pipeline at test time.

**VI-reID.** VI-reID has recently been explored compared to single-modality reID, according to the wide spread of RGB-IR cameras. VI-reID methods focus on handling cross-modal discrepancies between RGB and IR images, while learning discriminative person representations. Early works try to learn discriminative features generalizable across different modalities. They adopt classification and/or triplet losses, widely used in single-modality reID methods [36, 40], which however do not mitigate the cross-modal discrepancies explicitly. To address this problem, recent methods use a cross-modal triplet loss, where positive/negative pairs and an anchor are sampled from person images with different modalities [7, 37, 41]. For example, RGB images are used as anchors, while exploiting IR ones as positive/negative samples. These approaches encourage the features, obtained from person images of the same identity but having different modalities, to be similar, providing person representations robust to the cross-modal discrepancies. More recently, DDAG [39] proposes to leverage a graph attention network in order to consider cross-modal relations between RGB and IR images explicitly. VI-reID methods based on generative adversarial networks (GANs) alleviate the cross-modal discrepancies in an image level. For example, they synthesize novel IR person images, with an identity-preserving constraint [34] or cycle consistency [35], given RGB inputs, in order to compare person images with the same modality. Other approaches to leveraging adversarial learning techniques for VI-reID are to disentangle identity-related features from person representations [3], or to exploit a modality discriminator to better align feature distributions of RGB/IR images [4]. Although GANs better capture discriminative factors for person reID, they require lots of parameters and heuristics to train networks [28]. In contrast to current VI-reID methods, we address the cross-modal discrepancies in a pixel level. To this end, we align semantically related regions explicitly via dense cross-modal correspondences, which also allows discriminative feature learning, even from misaligned person images.

**Cross-modal image retrieval.** VI-reID is closely related to cross-modal image retrieval that focuses on finding matches between images of different modalities, *e.g.*, sketch/natural images [27, 29], and RGB/IR images [1, 20]. Existing works typically employ a siamese network [42] to learn a metric function between input image pairs [1, 29], or disentangle feature representations into modality shared- and specific- embeddings [20, 27]. They attempt to alleviate the cross-modal discrepancies between multi-modal images in an image-level. We instead address the discrepancies in a pixel-level by leveraging dense correspondences.

**Correspondence.** Establishing correspondences between images has long been of particular importance in many computer vision tasks, including depth prediction [13, 43], optical flow [2, 6], 3D scene reconstruction [16, 51], and colorization [11, 44]. In context of person reID, the work of [31] leverages dense correspondences to learn a metric function for single-modality person reID. The learned metric function, however, is required even at test time, demanding large computational power and memory. In contrast, we exploit the correspondences as explicit regularizers to guide the feature learning during training only, enabling a simple cosine distance computation between person representations at test time.

## 3. Approach

We describe in this section an overview of our framework for VI-reID (Sec. 3.1), and present detailed descriptions of a network architecture (Sec. 3.2) and training losses (Sec. 3.3).

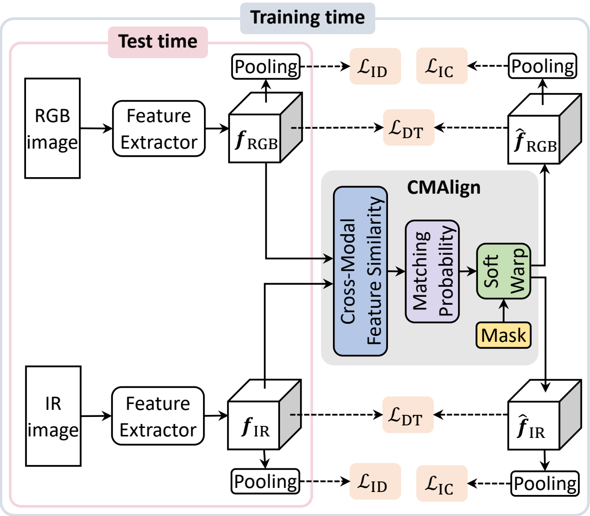

Training time
Test time
RGB image
Feature Extractor
$\mathbf{f}_{\mathrm{RGB}}$
Pooling
$\mathcal{L}_{\mathrm{ID}}$
$\mathcal{L}_{\mathrm{IC}}$
Pooling
$\mathcal{L}_{\mathrm{DT}}$
$\hat{\mathbf{f}}_{\mathrm{RGB}}$
CMAlign
Cross-Modal Feature Similarity
Matching Probability
Soft Warp
Mask
IR image
Feature Extractor
$\mathbf{f}_{\mathrm{IR}}$
$\mathcal{L}_{\mathrm{DT}}$
$\hat{\mathbf{f}}_{\mathrm{IR}}$
Pooling
$\mathcal{L}_{\mathrm{ID}}$
$\mathcal{L}_{\mathrm{IC}}$
Pooling

Figure 2: Overview of our framework for VI-reID. We extract RGB and IR features, denoted by $\mathbf{f}_{\mathrm{RGB}}$ and $\mathbf{f}_{\mathrm{IR}}$, respectively, using a two-stream CNN. The CMAlign module computes cross-modal feature similarities and matching probabilities between these features, and aligns the cross-modal features w.r.t each other using soft warping, together with parameter-free person masks to mitigate ambiguous matches between background regions. We exploit both original RGB and IR features and aligned ones ($\hat{\mathbf{f}}_{\mathrm{RGB}}$ and $\hat{\mathbf{f}}_{\mathrm{IR}}$) during training, and incorporate them into our objective function consisting of ID ($\mathcal{L}_{\mathrm{ID}}$), ID consistency ($\mathcal{L}_{\mathrm{IC}}$) and dense triplet ($\mathcal{L}_{\mathrm{DT}}$) terms. At test time, we compute cosine distances between person representations, obtained by pooling RGB and IR features. See text for details.

### 3.1. Overview

We show in Fig. 2 an overview of our framework for VI-reID. We first extract RGB and IR features from corresponding person images, and then align the features with a CMAlign module. It establishes dense cross-modal correspondences between RGB and IR features, and warps these features w.r.t each other using corresponding matching probabilities. Note that we exploit the CMAlign module at training time only, enabling an efficient inference at test time. To train our framework, we exploit three terms: ID ($\mathcal{L}_{\mathrm{ID}}$), ID consistency ($\mathcal{L}_{\mathrm{IC}}$), and dense triplet ($\mathcal{L}_{\mathrm{DT}}$) losses. The ID loss applies to each feature from RGB or IR images, separately, similar to single-modality reID [12]. It enforces the features from person images of the same identity to be the same, while providing different ones for the images of different identities. The ID consistency and dense triplet terms exploit the matching probabilities, and encourage RGB and IR features from the same identity to reconstruct one another in a pixel-level, while those from different identities do not. The person representations obtained using these terms are thus robust to the cross-modal discrepancies between RGB and IR images. Note that we use identification labels alone to train our model, without exploiting auxiliary supervisory signals, such as *e.g.*, body parts [15] or landmarks [23], for the alignment. Note also that all components in our model are fully differentiable, making it possible to train the whole network end-to-end.

### 3.2. Network Architecture

**Feature extractor.** We use a two-stream CNN to extract feature maps of size $h \times w \times d$ from a pair of RGB/IR person images, where $h$, $w$ and $d$ are the height, width and number of channels, respectively. Assuming that the cross-modal discrepancies between RGB/IR images mainly lie in low-level features [36, 40], we use separate parameters specific to the input modalities for shallow layers, while sharing the remaining ones for others.

**CMAlign.** The CMAlign module aligns RGB and IR features bidirectionally, *i.e.*, from RGB to IR and from IR to RGB, using dense cross-modal correspondences in a probabilistic way. In the following, we describe an IR-to-RGB alignment. The other case can be performed similarly.

For the IR-to-RGB alignment, we compute local similarities between all pairs of RGB and IR features. Concretely, we compute cosine similarities between RGB and IR features, denoted by $\mathbf{f}_{\mathrm{RGB}} \in \mathbb{R}^{h\times w\times d}$ and $\mathbf{f}_{\mathrm{IR}} \in \mathbb{R}^{h\times w\times d}$, respectively, as follows:

$$C(\mathbf{p},\mathbf{q}) = \frac{\mathbf{f}_{\mathrm{RGB}}(\mathbf{p})^{\top}\mathbf{f}_{\mathrm{IR}}(\mathbf{q})}{\|\mathbf{f}_{\mathrm{RGB}}(\mathbf{p})\|_2\|\mathbf{f}_{\mathrm{IR}}(\mathbf{q})\|_2}, \tag{1}$$

where $\|\cdot\|_2$ computes the L2 norm of a vector. We denote by $\mathbf{f}_{\mathrm{RGB}}(\mathbf{p})$ and $\mathbf{f}_{\mathrm{IR}}(\mathbf{q})$ RGB and IR features of size $d$ at position $\mathbf{p}$ and $\mathbf{q}$, respectively. Based on the similarities, we compute RGB-to-IR matching probabilities using a softmax function as follows:

$$P(\mathbf{p},\mathbf{q}) = \frac{\exp(\beta C(\mathbf{p},\mathbf{q}))}{\sum_{\mathbf{q}'}\exp(\beta C(\mathbf{p},\mathbf{q}'))}, \tag{2}$$

where we denote by $P$ a matching probability, a 4D tensor of size $h \times w \times h \times w$, and $\beta$ is a temperature parameter. Note that we can establish dense correspondences explicitly from RGB to IR images, by applying an argmax operator to the matching probabilities for each RGB feature, *i.e.*, $\mathrm{argmax}_{\mathbf{q}} P(\mathbf{p},\mathbf{q})$. This offers reliable cross-modal correspondences for semantically similar regions, but aligning IR and RGB features using the hard correspondences is problematic. The correspondences are easily distracted by background clutter and image-specific details (*e.g.*, texture and occlusion), and appearance variations between RGB and IR images are even more significant. Moreover, we could not establish correspondences between different background regions, *e.g.*, from person images captured with different surrounding environments. To alleviate these problems, we instead exploit the matching probabilities, and align IR and RGB features between foreground regions only, typically correspond to persons, via soft warping as follows:

$$\begin{aligned}\hat{\mathbf{f}}_{\mathrm{RGB}}(\mathbf{p}) &= \\ &M_{\mathrm{RGB}}(\mathbf{p})\mathcal{W}(\mathbf{f}_{\mathrm{IR}}(\mathbf{p})) + (1 - M_{\mathrm{RGB}}(\mathbf{p}))\mathbf{f}_{\mathrm{RGB}}(\mathbf{p}),\end{aligned} \tag{3}$$

where we denote by $\hat{\mathbf{f}}_{\mathrm{RGB}} \in \mathbb{R}^{h\times w\times d}$ and $M_{\mathrm{RGB}} \in \mathbb{R}^{h\times w}$ a reconstructed RGB feature by the IR-to-RGB alignment and a person mask, respectively. We denote by $\mathcal{W}$ a soft warping operator that aggregates features using the matching probabilities, defined as follows:

$$\mathcal{W}(\mathbf{f}_{\mathrm{IR}}(\mathbf{p})) = \sum_{\mathbf{q}} P(\mathbf{p},\mathbf{q})\mathbf{f}_{\mathrm{IR}}(\mathbf{q}). \tag{4}$$

The person mask ensures that the features $\hat{\mathbf{f}}_{\mathrm{RGB}}$, for person regions are reconstructed by aggregating IR features in a probabilistic way, while others come from original RGB features $\mathbf{f}_{\mathrm{RGB}}$. This reconstruction together with ID consistency and dense triplet losses encourages our model to provide similar person representations, regardless of image modalities, for the corresponding regions. To infer the mask without ground-truth labels, we assume that features, learned with ID labels for the reID task, are highly activated on person regions than other parts, and compute an activation map based on L2 norms of the local feature vectors, denoted by $\mathbf{g}_{\mathrm{RGB}} \in \mathbb{R}^{h\times w}$ for an RGB feature, as follows:

$$\mathbf{g}_{\mathrm{RGB}}(\mathbf{p}) = \|\mathbf{f}_{\mathrm{RGB}}(\mathbf{p})\|_2. \tag{5}$$

With the activation map of an RGB feature, $\mathbf{g}_{\mathrm{RGB}}$, at hand, we define a person mask for an RGB feature as follows:

$$M_{\mathrm{RGB}} = f(\mathbf{g}_{\mathrm{RGB}}), \tag{6}$$

where $f$ performs min-max normalization:

$$f(\mathbf{x}) = \frac{\mathbf{x} - \min(\mathbf{x})}{\max(\mathbf{x}) - \min(\mathbf{x})}. \tag{7}$$

The CMAlign, which is a non-parametric module that operates directly on the features obtained from the feature extractor, facilitates learning robust person representations by providing the following advantages in VI-reID: First, a cross-modal alignment helps to alleviate the discrepancies between RGB and IR images in a pixel-level, allowing to suppress modality-related features from person representations more effectively, even with misaligned person images; Second, a dense alignment allows our network to focus on learning local features, especially for person regions, further enhancing the discriminative power of person representations. Note that a pair of RGB and IR images does not have to be of the same identity in our framework, enabling exploiting both positive and negative pairs for training.

### 3.3. Loss

We exploit ground-truth ID labels of person images to train our model with an overall objective function as follows:

$$\mathcal{L} = \mathcal{L}_{\mathrm{ID}} + \lambda_{\mathrm{IC}}\mathcal{L}_{\mathrm{IC}} + \lambda_{\mathrm{DT}}\mathcal{L}_{\mathrm{DT}}, \tag{8}$$

where $\mathcal{L}_{\mathrm{ID}}$, $\mathcal{L}_{\mathrm{IC}}$ and $\mathcal{L}_{\mathrm{DT}}$ are ID, ID consistency, and dense triplet losses, respectively. $\lambda_{\mathrm{IC}}$ and $\lambda_{\mathrm{DT}}$ are hyperparameters to balance corresponding terms. In the following, we present a detailed description of each term in the loss.

**ID loss ($\mathcal{L}_{\mathrm{ID}}$).** As an ID loss, we adopt a sum of classification and hard triplet losses [12] using image-level person representations, which have shown the effectiveness on learning discriminative person features in single-modality person reID. We denote by $\phi(\mathbf{f}_{\mathrm{RGB}}) \in \mathbb{R}^d$ and $\phi(\mathbf{f}_{\mathrm{IR}}) \in \mathbb{R}^d$ image-level person representations for the RGB and IR features, respectively, which are obtained by applying a GeM pooling operation [26] to each feature. To compute the classification term, we feed each image-level feature, $\phi(\mathbf{f}_{\mathrm{RGB}})$ and $\phi(\mathbf{f}_{\mathrm{IR}})$, into a same classifier to predict class probabilities, that is, likelihoods of being particular identities for the image-level feature, where the classifier consists of a Batch Normalization layer [14], followed by a fully-connected layer with a softmax activation [22]. We then compute a cross-entropy between the class probabilities and ground-truth identities. The hard triplet term is also computed using image-level person representations, obtained from anchor, positive, and negative images, where the anchor and positive ones share the same ID label, while other pairs do not. Note that the ID loss does not address the cross-modal discrepancies between RGB and IR images explicitly.

**ID consistency loss ($\mathcal{L}_{\mathrm{IC}}$).** We design a term to consider the cross-modal discrepancies between RGB and IR features in an image-level. Suppose that we have a positive pair with the same identity but having different modalities, *i.e.*, RGB and IR images are of the same identity. The features $\hat{\mathbf{f}}_{\mathrm{RGB}}$ for person regions are reconstructed by aggregating IR features $\mathbf{f}_{\mathrm{IR}}$, suggesting that the identity of the reconstruction $\hat{\mathbf{f}}_{\mathrm{RGB}}$ should be the same as ground-truth identity for the original features $\mathbf{f}_{\mathrm{RGB}}$ and $\mathbf{f}_{\mathrm{IR}}$. More specifically, image-level representations of $\phi(\hat{\mathbf{f}}_{\mathrm{RGB}})$ and $\phi(\hat{\mathbf{f}}_{\mathrm{IR}})$ should have the same ID labels as corresponding positive counterparts of different modalities, $\mathbf{f}_{\mathrm{IR}}$ and $\mathbf{f}_{\mathrm{RGB}}$, respectively. To implement this idea, we define an ID consistency loss as a cross-entropy using image-level representations, similar to the classification term in the ID loss. We instead exploit reconstructed features, $\phi(\hat{\mathbf{f}}_{\mathrm{RGB}})$ and $\phi(\hat{\mathbf{f}}_{\mathrm{IR}})$. Note that we use the same classifier as in the ID loss. The ID consistency loss enforces ID predictions from person images of the same identity but with different modalities to be consistent, allowing to suppress modality-related features from person representations. Moreover, the reconstructions, $\phi(\hat{\mathbf{f}}_{\mathrm{RGB}})$ and $\phi(\hat{\mathbf{f}}_{\mathrm{IR}})$ provide an effect of offering additional samples to train the classifier, further guiding the discriminative person representation learning.

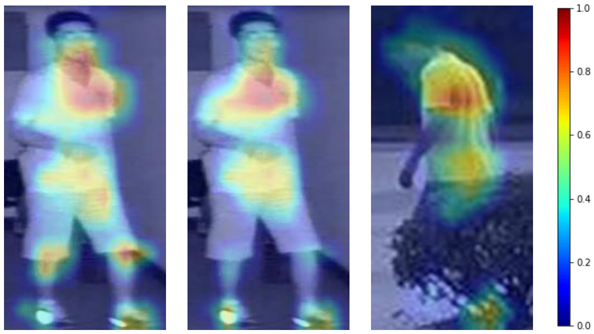


Figure 3: Visualization of person masks for IR and RGB images, $M_{\mathrm{IR}}$(left) and $M_{\mathrm{RGB}}$(right), and a corresponding co-attention map, $A_{\mathrm{IR}}$(middle). We overlay the masks and the attention map over corresponding images from SYSU-MM01 [36]. We can see that the IR image depicts a person with fully visible body parts, whereas the person of the same identity in the RGB image is partially occluded (lower body). The co-attention map highlights image regions that are mutually visible in both images and suppresses others using dense cross-modal alignments via soft warping. (Best viewed in color.)

**Dense triplet loss ($\mathcal{L}_{\mathrm{DT}}$).** The ID loss facilitates learning discriminative person representations, and the ID consistency term alleviates the cross-modal discrepancies explicitly. They, however, focus on learning image-level person representations, which prohibits discriminative feature learning, especially when the person images are occluded or misaligned. To address this problem, we introduce a dense triplet loss. It locally compares original features and reconstructed ones using the features of different modalities, encouraging final image-level person representations to be discriminative, while alleviating the cross-modal discrepancies in a pixel-level. A straight-forward approach is to compute L2 distances between local features, which is, however, suboptimal in that this does not take occluded regions into consideration. This is particularly problematic when each of person images in a pair depicts disassociated human parts. Enforcing local alignments between the entire person regions in this case is infeasible, and maybe even harmful. To circumvent this issue, we incorporate a co-attention map highlighting person regions visible in both RGB and IR images. This considers feature alignments within mutually visible foreground regions only to compute the dense triplet loss. We define a co-attention map, denoted by $A_{\mathrm{RGB}} \in \mathbb{R}^{h \times w}$ for an RGB image, as follows:

$$A_{\mathrm{RGB}}(\mathbf{p}) = M_{\mathrm{RGB}}(\mathbf{p})\mathcal{W}(M_{\mathrm{IR}}(\mathbf{p})). \tag{9}$$

For $\mathcal{W}(M_{\mathrm{IR}}(\mathbf{p}))$, in this case, we compute the matching probabilities $P$ between $\mathbf{f}_{\mathrm{RGB}}$ and $\mathbf{f}_{\mathrm{IR}}$, similar to (4), whereas the person masks are exploited for soft warping. Namely, the co-attention map $A_{\mathrm{RGB}}$ is an intersection of the RGB person mask $M_{\mathrm{RGB}}(\mathbf{p})$ and the warped IR one, w.r.t the RGB image, $\mathcal{W}(M_{\mathrm{IR}}(\mathbf{p}))$. We compute a co-attention map for an IR image similarly, and show in Fig. 3 an example of a co-attention map. Note that we define co-attention maps between positive pairs of the same identity only. Note also that we perform min-max normalization $f$ on the obtained co-attention map, which we omit for notational brevity.

To facilitate training with the dense triplet term, we sample a triplet of anchor, positive, and negative images, where the anchor and other two images have different modalities *e.g.*, an RGB image for the anchor, and IR images for a pair of positive and negative samples. We use the superscripts, $\mathrm{a}$, $\mathrm{p}$, and $\mathrm{n}$ to indicate features from anchor, positive, and negative images, respectively. For example, we denote by $\hat{\mathbf{f}}^{\mathrm{p}}_{\mathrm{RGB}}$ a reconstructed RGB feature using an anchor $\mathbf{f}^{\mathrm{a}}_{\mathrm{RGB}}$ and a positive pair $\mathbf{f}^{\mathrm{p}}_{\mathrm{IR}}$ with the same identity as the anchor $\mathbf{f}^{\mathrm{a}}_{\mathrm{RGB}}$. Similarly, $\hat{\mathbf{f}}^{\mathrm{n}}_{\mathrm{IR}}$ is a reconstructed IR feature using an anchor $\mathbf{f}^{\mathrm{a}}_{\mathrm{IR}}$ and a negative pair $\mathbf{f}^{\mathrm{n}}_{\mathrm{RGB}}$ with the different identity from the anchor $\mathbf{f}^{\mathrm{a}}_{\mathrm{IR}}$. With co-attention maps at hand, we define the dense triplet loss as follows:

$$\mathcal{L}_{\mathrm{DT}} = \sum_{i\in\{\mathrm{RGB},\mathrm{IR}\}} \sum_{\mathbf{p}} A_i(\mathbf{p})[d_i^+(\mathbf{p}) - d_i^-(\mathbf{p}) + \alpha]_+, \quad (10)$$

where $\alpha$ is a pre-defined margin and the operation $[\cdot]_+$ indicates $\max(0,\cdot)$. $d_i^+(\mathbf{p})$ and $d_i^-(\mathbf{p})$ compute local distances between an anchor feature and reconstructed ones from positive and negative images, respectively, as follows:

$$d_i^+(\mathbf{p}) = \|\mathbf{f}_i^{\mathrm{a}}(\mathbf{p}) - \hat{\mathbf{f}}_i^{\mathrm{p}}(\mathbf{p})\|_2, d_i^-(\mathbf{p}) = \|\mathbf{f}_i^{\mathrm{a}}(\mathbf{p}) - \hat{\mathbf{f}}_i^{\mathrm{n}}(\mathbf{p})\|_2. \quad (11)$$

Note that the reconstructions, $\hat{\mathbf{f}}_i^{\mathrm{p}}$ and $\hat{\mathbf{f}}_i^{\mathrm{n}}$, are the aggregations of similar features w.r.t the anchor $\mathbf{f}_i^{\mathrm{a}}$ from positive and negative images, respectively. We can thus interpret that our loss enforces an aggregation of similar features from negative images to be distant in the embedding space, compared to its positive counterpart by a margin. This is similar to the typical triplet loss [12, 30], but ours penalizes incorrect distances for all local features visible in both anchor and positive images in a soft manner. Note that this local association is possible due to the CMAlign module that performs the dense cross-modal alignment between RGB and IR person images in a probabilistic way.

# 4. Experiments

In this section, we present a detailed analysis and evaluation of our approach including ablation studies on different losses and network architectures.

## 4.1. Implementation details

**Dataset.** We use two benchmarks for evaluation: 1) The RegDB dataset [24] contains 412 persons, where each person has 10 visible and 10 far-infrared images collected by dual camera systems. Following the experimental protocol in [24], we divide the dataset into training and test splits randomly, each of which includes non-overlapping 206 identities. We test our model in both visible-to-IR and IR-to-visible settings, which correspond to retrieving IR images from RGB ones and RGB images from IR ones, respectively, and report the results averaged over 10 trials with different training/test splits. 2) SYSU-MM01 [36] is a large-scale dataset for VI-reID, consisting of RGB and IR images obtained by four visible and two near-infrared sensors, respectively. Concretely, it contains 22,258 visible and 11,909 near-infrared images with 395 identities for training. The test set contains 96 identities with 3,803 near-infrared images for a query set and 301 visible images for a gallery set. We adopt the evaluation protocol in [36], which uses all-search and indoor-search modes for testing, where the gallery sets for the former and the latter contain images captured by all four and two indoor visible cameras, respectively. Note that all our results are obtained by taking an average value over 4 training and test runs.

**Training.** Following the previous VI-reID methods [3, 21, 39], we adopt ResNet50 [10], trained for ImageNet classification [5], as our backbone network. The backbone networks for visible and infrared images share the parameters, except for the first residual blocks that take images of different modalities, and the stride of the last convolutional block is set to 1. We resize each person image to the size of 288 × 144, and apply horizontal flipping for data augmentation. We set the size of a person representation $d$ to 2,048. For a mini-batch, we randomly choose 8 identities from each modality and sample 4 person images for each identity. We train our model for 80 epochs with a batch size of 64, using the SGD optimizer with momentum of 0.9 and weight decay of 5e-4. We use a warm-up strategy [22], gradually raising learning rates for the backbones and other parts of the network up to 1e-2 and 1e-1, respectively, which are then decayed by a factor of 10 at the 20th and 50th epochs. We use a grid search to set hyper-parameters: $\lambda_{\mathrm{IC}} = 1$, $\lambda_{\mathrm{DT}} = 0.5$, $\alpha = 0.3$, $\beta = 50$. Note that we employ BNN trick [22] during training only, exploiting ResNet50 [10] at test time without any additional parameters. We implement our model with `PyTorch` [25] and train it end-to-end, taking about 6 and 8 hours for RegDB [24] and SYSU-MM01 [36], respectively, with a Geforce RTX 2080 Ti GPU.

## 4.2. Results

**Comparison with the state of the art.** We present in Table 1 a quantitative comparison of our method with the state of the art for VI-reID [3, 4, 17, 21, 34, 35, 36, 37, 38, 39]. We report mean average precision (mAP) (%) and rank-1 accuracy (%) for a single-shot setting on RegDB [24] and SYSU-MM01 [36]. From the table, we can see that our model sets a new state of the art for VI-reID, except for an indoor-search mode on SYSU-MM01 [36], where DDAG [39] shows better results. This method, however, requires additional parameters, other than the ResNet50 [10] backbone, for a feature refinement with self-attention at test

Table 1: Quantitative comparison with the state of the art for VI-reID. We measure mAP (%) and rank-1 accuracy (%) on the RegDB [24] and SYSU-MM01 [36] datasets and report the average and standard deviations over 4 training and test runs. Numbers in bold indicate the best performance and underscored ones indicate the second best.

| Methods | RegDB [24] | | | | SYSU-MM01 [36] | | | |
|---|---|---|---|---|---|---|---|---|
| | *Visible to Infrared* | | *Infrared to Visible* | | *All-search* | | *Indoor-search* | |
| | mAP | rank-1 | mAP | rank-1 | mAP | rank-1 | mAP | rank-1 |
| One-stream [36] | 14.02 | 13.11 | - | - | 13.67 | 12.04 | 22.95 | 16.94 |
| Two-stream [36] | 13.42 | 12.43 | - | - | 12.85 | 11.65 | 21.49 | 15.60 |
| Zero-Pad [36] | 18.90 | 17.75 | 17.82 | 16.63 | 15.95 | 14.80 | 26.92 | 20.58 |
| TONE [37] | 14.92 | 16.87 | - | - | 14.42 | 12.52 | 26.38 | 20.82 |
| HCML [37] | 20.08 | 24.44 | 22.24 | 21.70 | 16.16 | 14.32 | 30.08 | 24.52 |
| cmGAN [4] | - | - | - | - | 31.49 | 26.97 | 42.19 | 31.63 |
| BDTR [38] | 32.76 | 33.56 | 31.96 | 32.92 | 27.32 | 27.32 | 41.86 | 31.92 |
| D$^2$RL [35] | 44.10 | 43.40 | - | - | 29.20 | 28.90 | - | - |
| AlignGAN [34] | 53.60 | 57.90 | 53.40 | 56.30 | 40.70 | 42.40 | 54.30 | 45.90 |
| Xmodal [17] | 60.18 | 62.21 | 61.80 | 68.06 | 50.73 | 49.92 | - | - |
| Hi-CMD [3] | 66.04 | 70.93 | - | - | 35.94 | 34.94 | - | - |
| cm-SSFT [21] | 63.00 | 62.20 | - | - | 52.10 | 52.40 | - | - |
| DDAG [39] | 63.46 | 69.34 | 61.80 | 68.06 | 53.02 | 54.75 | **67.98** | **61.02** |
| Ours | **67.64** ± 0.08 | **74.17** ± 0.04 | **65.46** ± 0.18 | **72.43** ± 0.42 | **54.14** ± 0.33 | **55.41** ± 0.18 | 66.33 ± 1.27 | 58.46 ± 0.67 |

Table 2: Comparison of the average runtime for extracting a final person representation and the number of parameters required at test time.

| Methods | Model size (M) | | Runtime (ms) | |
|---|---|---|---|---|
| | RGB | IR | RGB | IR |
| AlignGAN [34] | 30.71 | 24.66 | 7.57 | 3.32 |
| Hi-CMD [3] | 52.63 | 52.63 | 4.43 | 4.43 |
| DDAG [39] | 40.32 | 40.32 | 2.03 | 2.03 |
| Ours | **23.52** | **23.52** | **1.90** | **1.90** |

Table 3: Quantitative comparison for variants of our model on the SYSU-MM01 dataset [36] (*All-search* mode).

| $\mathcal{L}_{\text{IC}}$ | $\mathcal{L}_{\text{DT}}$ | $A$ | Layer | mAP | rank-1 |
|---|---|---|---|---|---|
| ✗ | ✗ | - | - | 49.54 | 50.43 |
| ✓ | ✗ | - | 4, 5 | 52.88 | 54.44 |
| ✗ | ✓ | ✗ | 4, 5 | 50.08 | 50.38 |
| ✗ | ✓ | ✓ | 4, 5 | 51.23 | 51.06 |
| ✓ | ✓ | ✗ | 4, 5 | 52.78 | 53.44 |
| ✓ | ✓ | ✓ | 4 | 53.02 | 54.63 |
| ✓ | ✓ | ✓ | 5 | 53.81 | 54.66 |
| ✓ | ✓ | ✓ | 4, 5 | **54.14** | **55.41** |

time, while being outperformed by ours in other benchmarks. We can also see that our model achieves better results than cm-SSFT[1] [21] by a significant margin on both datasets. Note that cm-SSFT [21] uses multiple RGB and IR images to extract person representations, even at test time. That is, it exploits additional images of different modalities, *e.g.*, multiple IR images to extract features from an RGB input. cm-SSFT [21] is thus computationally expensive, and requires a lot of memory. Overall, the experimental results on the standard benchmarks demonstrate that our approach provides person representations robust to the cross-modal discrepancies and intra-class variations across RGB and IR images. Qualitative comparisons along with rank-10 accuracy (%) can be found in the supplementary material.

**Parameter and runtime analysis.** We compare in Table 2 the average runtime to extract a final person representation. For fair comparison, we measure the average runtime over 50 executions, for person images of the size 288 × 144 on the same machine with a Geforce RTX 2080 Ti GPU. Table 2 also compares the number of network parameters required at test time. Our method is fastest among the state of the art, and uses the smallest number of parameters, as it does not use any additional parameters, except the ones for a backbone network, at test time. Other methods on the contrary exploit additional layers or networks.

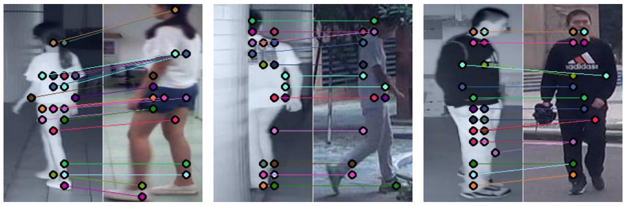
### 4.3. Discussion

**Ablation study.** We show in Table 3 an ablation analysis on training losses and the CMAlign module. We train variants of our model using different combinations of loss terms, $\mathcal{L}_{\text{IC}}$ and $\mathcal{L}_{\text{DT}}$, and co-attention map $A$, while adding CMAlign modules to different layers of the backbone network. We compare the performance in terms of mAP and rank-1 accuracy on SYSU-MM01 [36] under the *all-search* mode. For the baseline model in the first row, we exclude the CMAlign module and train it using the ID loss alone. Overall, we can see that the baseline shows the worst performance, indicating that incorporating the CMAlign module is beneficial for VI-reID. For example, exploiting the CMAlign module with either the ID consistency term (the second

[1] For cm-SSFT [21], we report in Table 1 the results obtained without using a random erasing technique [50] and a BNN trick [22], similar to ours, for fair comparison. The results are taken from Table 4 of [21].

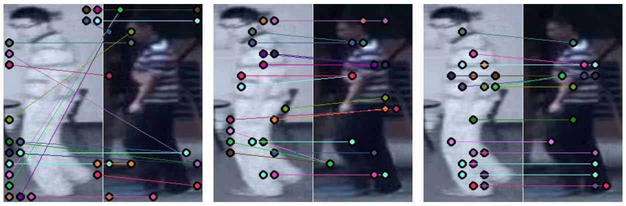
row) or the dense triplet term coupled with the co-attention map (the fourth row) boosts the performance significantly. This is because the ID consistency term mainly addresses cross-modal discrepancies in an image-level and the dense triplet term handles them in a pixel-level, while further enhancing the discriminative power of person representations. From the second, fourth, and last rows, we can observe that using all losses and the co-attention map gives the best results, suggesting that they are complementary to each other. Note that the co-attention map is particularly important for the dense triplet term, as shown in the fifth and last rows. Computing the loss on distractive regions (*e.g.*, occlusions and background clutter) may hinder learning discriminative representations. We also compare in the last three rows our models involving the CMAlign modules in different layers of the backbone network, where the modules are added on top of `conv4-6` and/or `conv5-3` of ResNet50 [10]. We can see that adding the modules to both `conv4-6` and `conv5-3` gives the best results, as this allows to consider cross-modal discrepancies in multiple levels of features.

**Visualizations of dense correspondences.** We show in Fig. 4 examples of cross-modal correspondences between RGB and IR images on SYSU-MM01 [36]. We can see that the matches are established well between persons of the same identity, and they are not influenced by cross-modal discrepancies, appearance variations, and background clutter. Specifically, our model provides local features that are robust against scale variations (Fig. 4(a)) and occlusions (Fig. 4(b)). This implies that our model is able to extract discriminative person representations with rich semantics, which are important for the person reID task, while alleviating the cross-modal discrepancies. In particular, our model offers local features that are robust to viewpoint variations (Fig. 4(c)), where a person's *sweatshirt* or *trousers* often matches to its pair regardless of front or side view. This indicates that our network provides local person features that are robust to viewpoint variations, which is particularly useful for VI-reID. This aspect of correspondences for reID is in contrast to typical correspondence tasks, *e.g.*, stereo matching and optical flow estimation, that favor viewpoint-specific matches. We also provide in Fig. 5 a visual comparison of correspondences for different configurations of losses. Our model trained with the ID loss alone is unable to establish reliable matches between cross-modal images and easily distracted by background clutter (Fig. 5(a)), mainly due to cross-modal discrepancies and a lack of discriminative power in local feature representations, particularly for person regions. The ID consistency loss handles the cross-modal discrepancies, establishing correspondences between local person representations from different modalities (Fig. 5(b)). The dense triplet loss further encourages each local feature to be discriminative, which in turn offers matching results focusing on person regions (Fig. 5(c)). The features trained by leveraging dense cross-modal correspondences are more discriminative, establishing matches focusing on person regions, while being robust to the cross-modal discrepancies. More examples can be found in the supplementary material.



(a) Scale variation (b) Occlusion (c) Viewpoint variation

Figure 4: Visualization of correspondences between RGB and IR images on SYSU-MM01 [36]. We show the top 20 matches chosen by matching probabilities. Our local person representations are robust to scale variations (a), occlusion (b), and viewpoint variations (c). (Best viewed in color.)



(a) $\mathcal{L}_{\text{ID}}$ (b) $\mathcal{L}_{\text{ID}} + \mathcal{L}_{\text{IC}}$ (c) $\mathcal{L}_{\text{ID}} + \mathcal{L}_{\text{IC}} + \mathcal{L}_{\text{DT}}$

Figure 5: Visual comparison of correspondences for different configurations of losses: (a) $\mathcal{L}_{\text{ID}}$; (b) $\mathcal{L}_{\text{ID}} + \mathcal{L}_{\text{IC}}$; (c) $\mathcal{L}_{\text{ID}} + \mathcal{L}_{\text{IC}} + \mathcal{L}_{\text{DT}}$. Our models in (b-c) exploit CMAlign modules. ID consistency and dense triplet terms help to alleviate the cross-modal discrepancies between RGB and IR images, while further enhancing the discriminative power of person features. (Best viewed in color.)

## 5. Conclusion

We have introduced a novel feature learning framework for VI-reID that exploits dense correspondences between cross-modal person images, allowing to learn person representations that are robust to intra-class variations and cross-modal discrepancies across RGB and IR person images. We have also proposed ID consistency and dense triplet losses exploiting pixel-level associations, enabling our model to learn more discriminative person representations. We set a new state of the art on standard benchmarks, outperforming other VI-reID methods by a significant margin. Extensive experimental results clearly demonstrate the effectiveness of our approach.

**Acknowledgements.** This research was partly supported by R&D program for Advanced Integrated-intelligence for Identification (AIID) through the National Research Foundation of KOREA (NRF) funded by Ministry of Science and ICT (NRF2018M3E3A1057289), Institute for Information and Communications Technology Promotion (IITP) funded by the Korean Government (MSIP) under Grant 2016-0-00197, and Yonsei University Research Fund of 2021 (2021-22-0001).

## References

[1] Cristhian A Aguilera, Francisco J Aguilera, Angel D Sappa, Cristhian Aguilera, and Ricardo Toledo. Learning cross-spectral similarity measures with deep convolutional neural networks. In *CVPR Workshop*, 2016. 3

[2] Thomas Brox and Jitendra Malik. Large displacement optical flow: Descriptor matching in variational motion estimation. *IEEE TPAMI*, 2010. 3

[3] Seokeon Choi, Sumin Lee, Youngeun Kim, Taekyung Kim, and Changick Kim. Hi-CMD: Hierarchical cross-modality disentanglement for visible-infrared person re-identification. In *CVPR*, 2020. 1, 3, 6, 7

[4] Pingyang Dai, Rongrong Ji, Haibin Wang, Qiong Wu, and Yuyu Huang. Cross-modality person re-identification with generative adversarial training. In *IJCAI*, 2018. 1, 3, 6, 7

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009. 6

[6] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. FlowNet: Learning optical flow with convolutional networks. In *ICCV*, 2015. 3

[7] Zhanxiang Feng, Jianhuang Lai, and Xiaohua Xie. Learning modality-specific representations for visible-infrared person re-identification. *IEEE TIP*, 2019. 1, 2

[8] Yang Fu, Yunchao Wei, Yuqian Zhou, Honghui Shi, Gao Huang, Xinchao Wang, Zhiqiang Yao, and Thomas Huang. Horizontal pyramid matching for person re-identification. In *AAAI*, 2019. 2

[9] Yixiao Ge, Zhuowan Li, Haiyu Zhao, Guojun Yin, Shuai Yi, Xiaogang Wang, et al. FD-GAN: Pose-guided feature distilling GAN for robust person re-identification. In *NeurIPS*, 2018. 2

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6, 8

[11] Mingming He, Dongdong Chen, Jing Liao, Pedro V Sander, and Lu Yuan. Deep exemplar-based colorization. *ACM TOG*, 2018. 3

[12] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 2, 3, 5, 6

[13] Asmaa Hosni, Christoph Rhemann, Michael Bleyer, Carsten Rother, and Margrit Gelautz. Fast cost-volume filtering for visual correspondence and beyond. *IEEE TPAMI*, 2012. 3

[14] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015. 5

[15] Mahdi M Kalayeh, Emrah Basaran, Muhittin Gökmen, Mustafa E Kamasak, and Mubarak Shah. Human semantic parsing for person re-identification. In *CVPR*, 2018. 2, 4

[16] Angjoo Kanazawa, David W Jacobs, and Manmohan Chandraker. WarpNet: Weakly supervised matching for single-view reconstruction. In *CVPR*, 2016. 3

[17] Diangang Li, Xing Wei, Xiaopeng Hong, and Yihong Gong. Infrared-visible cross-modal person re-identification with an x modality. In *AAAI*, 2020. 6, 7

[18] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *CVPR*, 2018. 2

[19] Yutian Lin, Liang Zheng, Zhedong Zheng, Yu Wu, Zhilan Hu, Chenggang Yan, and Yi Yang. Improving person re-identification by attribute and identity learning. *PR*, 2019. 2

[20] Fangcen Liu, Chenqiang Gao, Yongqing Sun, Yue Zhao, Feng Yang, Anyong Qin, and Deyu Meng. Infrared and visible cross-modal image retrieval through shared features. *IEEE TCSVT*, 2021. 3

[21] 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 *CVPR*, 2020. 6, 7

[22] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *CVPR Workshop*, 2019. 5, 6, 7

[23] Jiaxu Miao, Yu Wu, Ping Liu, Yuhang Ding, and Yi Yang. Pose-guided feature alignment for occluded person re-identification. In *ICCV*, 2019. 2, 4

[24] 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*, 2017. 6, 7

[25] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in PyTorch. 2017. 6

[26] Filip Radenović, Giorgos Tolias, and Ondřej Chum. Fine-tuning CNN image retrieval with no human annotation. *IEEE TPAMI*, 2018. 5

[27] Aneeshan Sain, Ayan Kumar Bhunia, Yongxin Yang, Tao Xiang, and Yi-Zhe Song. Stylemeup: Towards style-agnostic sketch-based image retrieval. In *CVPR*, 2021. 3

[28] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. In *NeurIPS*, 2016. 3

[29] Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. The sketchy database: learning to retrieve badly drawn bunnies. *ACM TOG*, 2016. 3

[30] Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A unified embedding for face recognition and clustering. In *CVPR*, 2015. 6

[31] Yantao Shen, Tong Xiao, Hongsheng Li, Shuai Yi, and Xiaogang Wang. End-to-end deep kronecker-product matching for person re-identification. In *CVPR*, 2018. 3

[32] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *ECCV*, 2018. 2

[33] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning discriminative features with multiple granularities for person re-identification. In *ACM MM*, 2018. 2

[34] 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 *ICCV*, 2019. 3, 6, 7

[35] Zhixiang Wang, Zheng Wang, Yinqiang Zheng, Yung-Yu Chuang, and Shin'ichi Satoh. Learning to reduce dual-level

discrepancy for infrared-visible person re-identification. In *CVPR*, 2019. 3, 6, 7

[36] Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong, and Jianhuang Lai. RGB-infrared cross-modality person re-identification. In *ICCV*, 2017. 1, 2, 4, 5, 6, 7, 8

[37] Mang Ye, Xiangyuan Lan, Jiawei Li, and Pong C Yuen. Hierarchical discriminative learning for visible thermal person re-identification. In *AAAI*, 2018. 1, 2, 6, 7

[38] Mang Ye, Xiangyuan Lan, Zheng Wang, and Pong C Yuen. Bi-directional center-constrained top-ranking for visible thermal person re-identification. *IEEE TIFS*, 2019. 6, 7

[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 *ECCV*, 2020. 1, 3, 6, 7

[40] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE TPAMI*, 2021. 1, 2, 4

[41] Mang Ye, Zheng Wang, Xiangyuan Lan, and Pong C Yuen. Visible thermal person re-identification via dual-constrained top-ranking. In *IJCAI*, 2018. 1, 2

[42] Sergey Zagoruyko and Nikos Komodakis. Learning to compare image patches via convolutional neural networks. In *CVPR*, 2015. 3

[43] Jure Žbontar and Yann LeCun. Stereo matching by training a convolutional neural network to compare image patches. *JMLR*, 2016. 3

[44] Bo Zhang, Mingming He, Jing Liao, Pedro V Sander, Lu Yuan, Amine Bermak, and Dong Chen. Deep exemplar-based video colorization. In *CVPR*, 2019. 3

[45] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. Densely semantically aligned person re-identification. In *CVPR*, 2019. 2

[46] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, Xin Jin, and Zhibo Chen. Relation-aware global attention for person re-identification. In *CVPR*, 2020. 2

[47] Liang Zheng, Yi Yang, and Alexander G Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016. 1

[48] Zhedong Zheng, Xiaodong Yang, Zhiding Yu, Liang Zheng, Yi Yang, and Jan Kautz. Joint discriminative and generative learning for person re-identification. In *CVPR*, 2019. 2

[49] Zhedong Zheng, Liang Zheng, and Yi Yang. A discriminatively learned CNN embedding for person re-identification. *ACM TOMM*, 2017. 2

[50] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *AAAI*, 2020. 7

[51] Tinghui Zhou, Yong Jae Lee, Stella X Yu, and Alyosha A Efros. FlowWeb: Joint image set alignment by weaving consistent, pixel-wise correspondences. In *CVPR*, 2015. 3

[52] Kuan Zhu, Haiyun Guo, Zhiwei Liu, Ming Tang, and Jinqiao Wang. Identity-guided human semantic parsing for person re-identification. In *ECCV*, 2020. 2