Modality-aware Style Adaptation for RGB-Infrared Person Re-Identification

Ziling Miao[1] , **Hong Liu**[1*] , **Wei Shi**[1] , **Wanlu Xu**[1] , **Hanrong Ye**[2]

[1]Key Laboratory of Machine Perception, Peking University, Shenzhen Graduate School, Shenzhen, China
[2]Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong

{zilingmiao, hongliu, pkusw, xuwanlu}@pku.edu.cn, hanrong.ye@connect.ust.hk

Abstract

RGB-infrared (IR) person re-identification is a challenging task due to the large modality gap between RGB and IR images. Many existing methods bridge the modality gap by style conversion, requiring high-similarity images exchanged by complex CNN structures, like GAN. In this paper, we propose a highly compact modality-aware style adaptation (MSA) framework, which aims to explore more potential relations between RGB and IR modalities by introducing new related modalities. Therefore, the attention is shifted from bridging to filling the modality gap with no requirement on high-quality generated images. To this end, we firstly propose a concise feature-free image generation structure to adapt the original modalities to two new styles that are compatible with both inputs by patch-based pixel redistribution. Secondly, we devise two image style quantification metrics to discriminate styles in image space using luminance and contrast. Thirdly, we design two image-level losses based on the quantified results to guide the style adaptation during an end-to-end four-modality collaborative learning process. Experimental results on two datasets SYSU-MM01 and RegDB show that MSA achieves significant improvements with little extra computation cost and outperforms the state-of-the-art methods.

1 Introduction

Person re-identification (Re-ID) aims to retrieve a target person across disjoint camera views [Ye *et al.*, 2020c; Ye *et al.*, 2020a]. Given probe images of a person-of-interest, Re-ID searches the gallery set to match images with the same identity. However, Re-ID based on RGB cameras usually fails to capture valid appearance information under poor lighting conditions. With the introduction of infrared cameras, the cross-modality RGB-Infrared (RGB-IR) person Re-ID is proposed to leverage the discrepancy of modalities originated from different imaging processes of RGB and IR cameras, where the probe consists of RGB images and the gallery consists of IR images or vice versa.

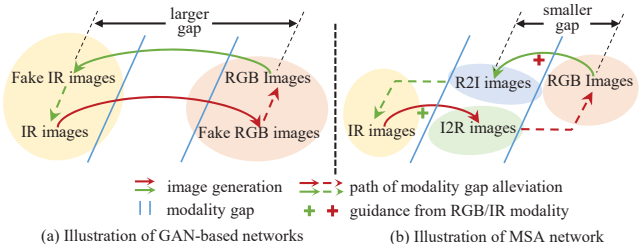

Figure 1: Illustration of the motivation of MSA. Different color blocks mean different styles. (a) The networks need to generate high-quality fake images to stride the large modality gap. (b) The MSA alleviates the modality gap from the view of increasing the input style diversity, which promotes the exploration of style-invariant features. So it faces a smaller gap of image generation without dependence on high-similarity generated images.

One critical challenge for RGB-IR person Re-ID is the large cross-modality gap between RGB and IR images. To reduce the modality gap, existing methods for RGB-IR Re-ID can be generally divided into two types. Some methods work in the feature space by designing some modality-specific and modality-shared embedding networks or classification layers [Dai *et al.*, 2018; Ye *et al.*, 2019b; Feng *et al.*, 2019]. However, these methods usually ignore the large gap existed in image space. The other methods apply GAN-based networks to achieve dual image style conversion while bridging the two modalities in image space [Wang *et al.*, 2019a; Wang *et al.*, 2019b], using high-similarity images exchanged with complex image generation structures, as shown in Fig. 1a.

To alleviate the modality gap, we propose a modality-aware style adaptation (MSA) framework to facilitate the mining of potential relations between RGB and IR modalities by introducing new related modalities. Different from existing methods, MSA enlightens image-to-image communication between RGB and IR modalities and adopts the knowledge to generate new modalities that are compatible with both RGB and IR modalities. As illustrated in Fig. 1b, MSA adapts RGB and IR images to new R2I and I2R styles in parallel under the guidance from the counterpart, filling the large modality gap utilizing connections among the four modalities. Firstly, a compact image generation framework

*Corresponding author.

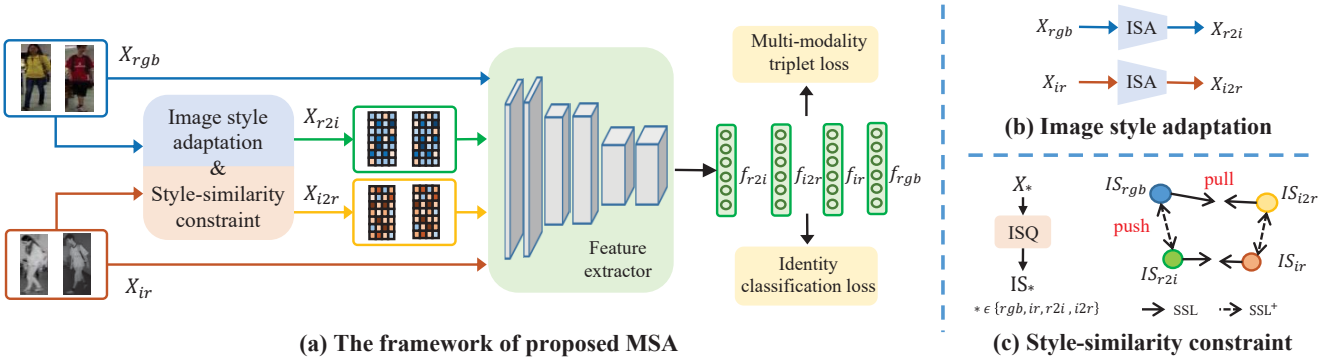

Figure 2: (a) The proposed MSA framework. In image sapce, the image style adaptation module (ISA) takes X_{rgb} and X_{ir} as input and outputs X_{r2i} and X_{i2r} under the constraint of the style similarity module (SSC), which is the combination of image style quantification (ISQ) and style similarity losses (SSL$^+$ and SSL). In feature space, images of the four modalities are fed into a common embedding network, then two losses are applied to the output embeddings. (b) ISA module with details in Section 3.1. (c) SSC module with details in Section 3.2. The network is trained end-to-end. (Please view in color.)

is designed without deep CNN structures since MSA does not require high-similarity generated images. In detail, MSA shuffles knowledge of patches in an original image and reconstructs constrained by two image-level losses, which are calculated by two devised image style quantification metrics from global and patch views. Subsequently, the two original and two generated modalities are fed into a common embedding network and features of the four modalities are constrained by identification loss and triplet loss. Above all, the trainable image generation structure and feature embedding network are optimized simultaneously by two image-level losses and two feature-level losses during an end-to-end training. The main contributions of this work are as follows:

- We propose a compact modality-aware style adaptation (MSA) framework, which introduces two new related modalities to help explore modality-invariant features.
- We design two image-level losses and combine them with feature-level losses to form a dual-space (image and feature space) total objective function.
- MSA outperforms the state-of-the-arts on two publicly available datasets. Specially, over 6% improvements of rank-1 and mAP on the RegDB dataset.

2 Related Works

2.1 RGB-Infrared Person Re-identification

For RGB-IR Re-ID, various methods are proposed to reduce the modality gap, and most of them can be divided into two types. The first type contains methods focusing on the feature space. For instance, [Wu *et al.*, 2017] researched three kinds of frameworks and used zero-padding to learn the common space for two different modalities. [Dai *et al.*, 2018] designed a cross-modality generative adversarial network (cmGAN) to reduce the modality gap by learning a highly integrated feature space. [Ye *et al.*, 2019a] designed two separate identity classifiers and a modality-sharable classifier to learn the discriminative information of different modalities. The second type aims at image generation, which reduces the modality gap by converting images from RGB modality to IR style or vice versa, while keeping the identity information. [Wang *et al.*, 2019b] reduced the image-level discrepancy through a dual image transformation. [Wang *et al.*, 2019a] generated fake IR images to learn more discriminative information of identities from different modalities.

Methods focusing on feature space are usually limited by the large image-level gap. However, existing image-level methods usually bridge the two modalities through deep CNN-based style conversion with a lot of extra parameters. MSA works in image space and design a compact two-layer image generation structure.

2.2 Image Similarity Metric

The GAN-based Re-ID methods use the feature–level discriminator to constrain the similarity of images. However, some image-level metrics comparing the images pixel by pixel for evaluation are still effective. For instance, [Wang *et al.*, 2004] proposed a structural similarity (SSIM) to compare two images from the view of luminance, contrast and structure. More generally, the peak signal-to-noise ratio (PSNR), which measures the image similarity by calculating the mean square error (MSE) of two images and the image-level L1 loss are widely used to measure the similarity of two images [Zhang *et al.*, 2019; Liu *et al.*, 2020]. Hardly existing RGB-IR Re-ID methods focus on style conversion in image space, therefore ignore to construct image-level distance metrics for style discrimination. In MSA, we fill this vacancy by designing two style similarity metrics used in the image space and further introduce two image-level losses.

3 The Proposed Method

MSA achieves style adaptation by reconstructing the source images while siphoning off knowledge from the target style, where the results are compatible with both source and target modalities. Based on that, we denote a triplet of images with form $\{X_s, X_a, X_t\}$ referring to images from the source, generated, and target modalities during style adaptation. And it can be adapted as $\{X_{rgb}, X_{r2i}, X_{ir}\}$ or $\{X_{ir}, X_{i2r}, X_{rgb}\}$ in

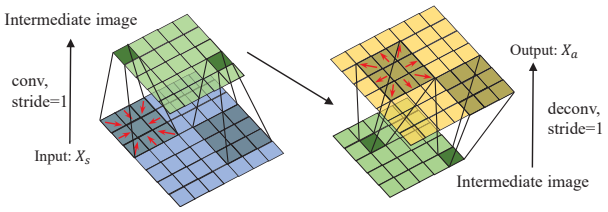

Figure 3: The process of ISA. It contains a convolution subprocess and a deconvolution subprocess to achieve pixel redistribution. Red arrows represent the direction of information flow.

MSA. The proposed framework is shown in Fig. 2a. In image space, MSA aims to generate two new modalities through dual style adaptation. In feature space, it focuses more on constructing a high integrated feature space where the cross-modality variations are alleviated.

3.1 Image Style Adaptation

The image style adaptation (ISA) module achieves the image reconstruction stage in style adaptation and designs a structure with concise operations. As shown in Fig. 3, firstly, based on the patch-level receptive field, the convolutional layer combines local pixel information of the input image into one pixel of the intermediate image. Secondly, the deconvolutional layer takes the intermediate image as input and redistributes the pixels with a learnable form in the output image. During training, two ISA modules are applied with the same structure but optimized separately. Above all, the process of ISA is defined as:

$$X_a = \boldsymbol{\mathcal{D}}(\boldsymbol{\mathcal{C}}(X_s)) \tag{1}$$

where $\boldsymbol{\mathcal{C}}$ and $\boldsymbol{\mathcal{D}}$ represent the convolution and deconvolution operations separately. X_s and X_a confirm the definition above. What worth noticed is there is no padding operation existed in ISA to avoid the loss of information during deconvolution. Moreover, the kernel sizes of $\boldsymbol{\mathcal{C}}$ and $\boldsymbol{\mathcal{D}}$ are both set as 3×3, with neglected influence on the image identity (the label) due to the small covering range.

3.2 Style Similarity Constraint

The style similarity constraint (SSC) module siphons off knowledge from the target style in style adaptation. As shown in Fig. 2c, SSC consists of an image style quantification (ISQ) block and two style similarity losses (SSL$^+$ and SSL). Specifically, it models the relations among modalities by discriminating the source, generated and target styles in image space (ISQ) and guides the style adaptation by calculating image-level style distances of them (SSL$^+$ and SSL).

Image style quantification. Firstly, we discriminate different styles by quantifying image style. Considering the patch-based ISA module, the image style quantification (ISQ) module is also proposed from a patch view using patch luminance and contrast of an image, which are described by the mean value μ and standard deviation σ [Wang *et al.*, 2004]. We denote an image $x \in \mathbb{R}^{C\times H\times W}$, where C represents the channel dimension, H and W represent the image size. As shown in Fig. 4, an $N \times N$ window conforming to the Gaussian

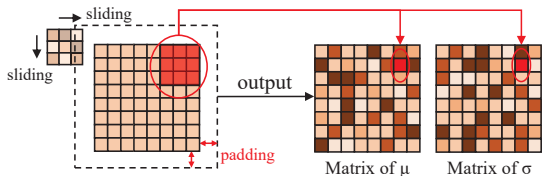

Figure 4: The process of ISQ. Here the values surrounded by the red circles in the output matrixes are calculated based on the pixels surrounded by the red circle in the initial image.

distribution is designed to separately calculate the μ and σ within the sliding window in each step. Subsequently, results in each step are concatenated keeping the topological relationship invariant. Based on that, the quantified image style $IS_x \in \mathbb{R}^{C\times H\times W}$ contains two parallel outputs:

$$\begin{aligned} IS_x &= \{IS_{x_\mu}, IS_{x_\sigma}\} \\ &= \{\lceil \sum_{i=1}^{W}\sum_{j=1}^{H} \phi(x_{i,j}) \rfloor, \lceil \sum_{i=1}^{W}\sum_{j=1}^{H} \sqrt{\phi({x^2}_{i,j})} \rfloor\} \end{aligned} \tag{2}$$

where

$$\phi(x_{i,j}) = G \odot \Gamma(x_{i,j}) \tag{3}$$

here, G represents a Gaussian filter and $\odot$ represents the element-wise product of matrixes. $\Gamma(\cdot)$ is a truncation function which takes the input pixel as center and cuts out a new $N \times N$ image. $x^2 \in \mathbb{R}^{C\times H\times W}$ is calculated by $x \odot x$. And $\lceil \rfloor$ is the concatenation operation. Especially, the IS_x is calculated parallel using Eq. 2 in each channel, and that of the whole image is the concatenation of C channels.

Style similarity loss. Secondly, after discriminating different styles using ISQ, two style similarity losses are designed to restrict the distances between the generated style and the source (target) style during style adaptation. As illustrated in Fig. 2c, the style distances between $\{X_t, X_a\}$ and $\{X_s, X_a\}$ are calculated differently. For the first pair, $\mathcal{L}_{SSL}$ is defined using the MSE loss:

$$\mathcal{L}_{SSL} = \frac{1}{C}(\sum_{c=1}^{C} \Delta IS^c_{x_\mu} + \sum_{c=1}^{C} \Delta IS^c_{x_\sigma}) \tag{4}$$

where

$$\Delta IS_{x_\mu} = (IS_{t_\mu} - IS_{a_\mu})^2 \tag{5}$$

and

$$\Delta IS_{x_\sigma} = {IS_{t_\sigma}}^2 + {IS_{a_\sigma}}^2 - 2 * IS_{(t\odot a)_\sigma} \tag{6}$$

here, c is the index of channel. Then, for $\{X_s, X_a\}$, which are highly similar to each other in terms of content, a more powerful global view constraint may work better [Zhang *et al.*, 2019]. Directly, the style of images is affected by the value of each pixel, thus a reinforced global-view image style $IS^+_x \in \mathbb{R}^{C\times H\times W}$ of image x is quantified as:

$$IS^+_x = x \tag{7}$$

and corresponding $\mathcal{L}_{SSL^+}$ is defined as:

$$\mathcal{L}_{SSL^+} = ||IS^+_s - IS^+_a||_1 \tag{8}$$

where $||\cdot||$ represents the l_1-$norm$ used to further sharpen the generated X_a compared with MSE [Zhang *et al.*, 2019]. More details about the function of $\mathcal{L}_{SSL^+}$ and $\mathcal{L}_{SSL}$ are discussed in Section 4.4.

Overall image-level loss. The style adaptation is mainly achieved in image space using the aforementioned modules. To propagate knowledge from both RGB and IR modalities to the new modalities in image space, MSA restricts the style distance of $\{X_t, X_a\}$ and $\{X_s, X_a\}$ simultaneously. Firstly, $\mathcal{L}_{SSL}$ is used to reduce the style distance between $\{X_t, X_a\}$ to force X_a to stay more closer to X_t in style. However, the effect of $\mathcal{L}_{SSL}$ is limited by the different content information of $\{X_t, X_a\}$. Moreover, due to the simple structure of ISA, the generated X_a is actually highly similar to X_s, which puts MSA in a dilemma where the similarity between $\{X_s, X_a\}$ is too high while that between $\{X_t, X_a\}$ is not high enough. Accordingly, the more powerful $\mathcal{L}_{SSL^+}$ is introduced to enlarge the style distance between $\{X_s, X_a\}$ on the one hand, and assist $\mathcal{L}_{SSL}$ to close $\{X_t, X_a\}$ on the other hand. As a result, the overall image-level loss is defined as:

$$\mathcal{L}_{img} = \alpha\mathcal{L}_{SSL} - \beta\mathcal{L}_{SSL^+} \tag{9}$$

3.3 Feature Representation Learning

We design a common embedding network $\boldsymbol{\mathcal{F}}$ for the four modalities as shown in Fig. 2a, to ensure that knowledge from one specific style can be broadcast to others through the back-propagation during end-to-end training. And the output embeddings are constrained by two feature-level losses.

3.4 Feature-level Loss

Firstly, identity (ID) classification loss [Liu *et al.*, 2018] is applied by optimizing the cross-entropy loss between labels and predicted possibilities. Secondly, an improved multi-modality triplet loss is developed as shown in Fig. 2c. Compared with the conventional one, it ignores the modality attribute of images and only leaves restrictions on the same identity for (a, p) and different identities for (a, n). Moreover, it can handles features from four modalities simultaneously. Above all, the multi-modality triplet loss with hard sample mining is defined as:

$$\begin{aligned}\mathcal{L}_{mmtri} = [\max D[\boldsymbol{\mathcal{F}}(x^i_{ma}), \boldsymbol{\mathcal{F}}(x^i_{mp})] - \\ \min D(\boldsymbol{\mathcal{F}}(x^i_{ma}), \boldsymbol{\mathcal{F}}(x^j_{mn})) + \lambda]_+,\end{aligned} \tag{10}$$

where λ is a margin parameter and $[z]_+ = max(z, 0)$. $D(\cdot)$ is the Euclidean distance. $(a, p, n) = (x^i_{ma}, x^i_{mp}, x^j_{mn})$, where i,j are unequal identities of persons. (ma, mp, mn) are corresponding modalities of the anchor and its hard samples, allowing any combination of modalities.

3.5 Multi-modality Collaborative Learning

The aforementioned modules are integrated by a multi-modality collaborative learning strategy and optimized by both image-level and feature-level losses. The overall loss function is formulated as:

$$\mathcal{L}_{total} = \mathcal{L}_{img} + \gamma\mathcal{L}_{mmtri} + L_{id}, \tag{11}$$

Algorithm 1 shows the overall training process of MSA. In conclusion, the trainable ISA is introduced to help search two 'new styles', with which the feature embedding network is optimal. Therefore, end-to-end training is necessary and it enables stronger losses from image and feature spaces.

Algorithm 1 Multi-modality collaborative learning of MSA

Input: RGB images:X_{rgb}, IR images: X_{ir}
Parameter: α, β, γ
Output: Configurations of MSA

```
Initialize the parameters of F, D and C
for each x_rgb, x_ir do
    Generate x_r2i, x_i2r using Eq. (1)
    for each triplet of {x_rgb, x_r2i, x_ir} do
        Calculate L_SSL1, L_SSL+1 using Eq. (4), (8).
    end for
    for each triplet of {x_ir, x_i2r, x_rgb} do
        Calculate L_SSL2, L_SSL+2 using Eq. (4), (8).
    end for
    L_img = α(L_SSL1 + L_SSL2) − β(L_SSL+1 + L_SSL+2)
    for each {x_rgb, x_ir, x_r2i, x_i2r} do
        Calculte F(rgb), F(ir), F(r2i), F(i2r)
        Calculte L_id
        Calculte L_mmtri using Eq. (10).
    end for
    L_total = L_img + γL_mmtri + L_id
end for
```

4 Experiments

4.1 Datasets and Settings

Datasets. (1) SYSU-MM01 [Wu *et al.*, 2017] is a large-scale and challenging dataset, containing 491 identities with 30,071 RGB images and 15,792 IR images. Among them, 395 persons with 22,580 RGB images and 11,909 IR images are divided into the training set, and 96 persons are divided into the testing set. During testing, we apply the more challenging *single-shot* mode. (2) RegDB [Nguyen *et al.*, 2017] contains 412 identities, which has 10 RGB images and 10 thermal images. The dataset is randomly divided into two halves, training and testing set following [Ye *et al.*, 2020c]. During testing, we adopt the *Visible2Thermal* mode which takes the RGB images as probe and IR images as gallery.

Implementation details. The ResNet-50 [He *et al.*, 2016] with pre-trained parameters on ImageNet [Krizhevsky *et al.*, 2012] is taken as our backbone, where the stride of the last convolutional layer is changed to one. During training, we randomly select six identities with four RGB images and four IR images sampled for each identity. We also adopt the random erasing [Lu *et al.*, 2020] for data augmentation. The Adam optimizer is used to guide the training process in 60 epochs. The λ in L_{mmtri} is set to 0.5. And the trade-off hyperparameters α, β and γ are set to 1:1:1 and 13:10:7 separately on SYSU-MM01 and RegDB datasets. During testing, only the RGB and IR images are taken as input of the feature embedding network.

4.2 Comparison with State-of-the-Art Methods

The proposed method is compared with the state-of-the-arts on two datasets, including Zero-Padding [Wu *et al.*, 2017], HCML [Ye *et al.*, 2018], cmGAN [Dai *et al.*, 2018], D-HSME [Hao *et al.*, 2019], D^{2}RL [Wang *et al.*, 2019b], AlignGAN [Wang *et al.*, 2019a], AT [Ye *et al.*, 2020a], Hi-CMD [Choi *et al.*, 2020], JSIA [Wang *et al.*, 2020], XIV [Li

Method	r = 1	r = 10	r = 20	mAP
Zero-Padding (ICCV17)	17.75	34.21	44.35	18.90
HCML (AAAI18)	24.44	47.53	56.78	20.08
D^{2}RL (CVPR19)	43.40	66.10	76.30	44.10
D-HSME (AAAI19)	50.85	73.36	81.66	47.00
AlignGAN (ICCV19)	57.90	-	-	53.60
XIV (AAAI20)	62.21	83.13	91.72	60.18
AT (TIP20)	69.60	-	-	69.84
Hi-CMD (CVPR20)	70.93	86.39	-	66.04
cm-SSFT (CVPR20)	72.30	-	-	72.90
SIM (IJCAI20)	74.47	-	-	75.29
Ours	**84.86**	**92.75**	**95.11**	**82.16**

Table 1: Comparison results(%) at Rank r with the state-of-the-art cross-modality Re-ID methods on the RegDB dataset.

Method	*All-search*		*Indoor-search*	
	r = 1	mAP	r = 1	mAP
Zero-Padding (ICCV17)	14.8	15.95	20.58	26.92
HCML (AAAI18)	14.32	16.16	24.52	30.08
D-HSME (AAAI19)	20.68	23.12	-	-
cmGAN (IJCAI18)	26.97	27.8	31.63	42.19
D^{2}RL (CVPR19)	28.9	29.20	-	-
AlignGAN (ICCV19)	42.4	40.7	45.9	54.3
Hi-CMD (CVPR20)	34.94	35.94	-	-
JSIA (AAAI20)	38.10	36.9	43.8	52.9
XIV (AAAI20)	49.92	50.73	-	-
cm-SSFT (CVPR20)	61.60	**63.20**	**70.50**	72.60
Ours	**63.13**	59.22	67.18	**72.74**

Table 2: Comparison results(%) at Rank r with the state-of-the-art cross-modality Re-ID methods on the SYSU-MM01 dataset.

et al., 2020], DDAG [Ye *et al.*, 2020b], SIM [Jia *et al.*, 2020] and cm-SSFT[Lu *et al.*, 2020]. Table 1 and Table 2 show the results over the RegDB and SYSU-MM01 datasets.

With a compact model, MSA outperforms the SOTA by 10.3% and 6.9% in terms of rank 1 and mAP on the RegDB dataset. On the SYSU-MM01 dataset, we reach the leading level from the above view, but outperform the SOTA in rank 10 and rank 20 accuracies which are not listed in the table, with improvements of 3.76% and 3.08% under *all-search* mode and 1.69% and 1.51% under *indoor-search* mode. Moreover, Compared with methods concerning the generation of new images, MSA does not rely on GAN-based components [Wang *et al.*, 2019a; Wang *et al.*, 2019b] and achieves not only channel-wise [Li *et al.*, 2020] but also pixel-wise image reconstruction. In addition, we align the settings with AGW [Ye *et al.*, 2020c] and DDAG [Ye *et al.*, 2020b] for a fair comparison as in Table 3. Results show that MSA is still effective without the improvement of backbone, data and network augmentation.

4.3 Ablation Study

Extensive ablation experiments are set to evaluate each component of MSA as shown in Table 4. To ensure the robustness of results, all experiments are performed using two backbones. Settings for each group are: 1) B: Baseline model trained with $\mathcal{L}_{id}$ and $\mathcal{L}_{mmtri}$ of original RGB and IR modalities like [Ye *et al.*, 2020c]. 2) B + ISA: Model trained with $\mathcal{L}_{id}$ and $\mathcal{L}_{mmtri}$ of at least three modalities, which is optimized only by feature-level losses. 3) B + ISA + SSC: Model trained with $\mathcal{L}_{total}$ of three modalities. 4) MSA: Model trained with $\mathcal{L}_{total}$ of four modalities.

Method	*All-search*		*Indoor-search*	
	r = 1	mAP	r = 1	mAP
AGW (TPAMI21)	47.50	47.65	54.17	62.97
DDAG (ECCV20)	54.75	**53.02**	61.02	67.98
Ours	**55.50**	52.57	**61.19**	**68.15**

Table 3: Comparison with AGW and DDAG using the same backbone, data and network augmentation on the SYSU-MM01 dataset.

Method	Modality		ResNet-50		DenseNet-121	
	R2I	I2R	r = 1	mAP	r = 1	mAP
B	×	×	42.83	41.97	31.96	30.02
B+ISA	✓	×	60.85	57.71	50.02	48.04
	×	✓	59.17	57.19	48.55	46.99
	✓	✓	62.08	58.80	51.91	49.80
B+ISA+SSC	✓	×	61.13	58.13	50.89	47.80
	×	✓	60.27	58.02	49.72	47.72
MSA	✓	✓	**63.13**	**59.22**	**54.13**	**52.03**

Table 4: Ablation study on the large-scale SYSU-MM01 dataset under the *all-search* mode.

Evaluation of R2I and I2R (ISA module). As shown in Table 4, the baseline does not perform well with the original modalities due to the large modality gap. MSA boosts the performance by introducing two new modalities, hence we evaluate if either of the R2I and I2R modalities works by comparing the 2nd group with the baseline. Results support the effectiveness of both new modalities generated by the designed ISA module and further verify the positive significance of generating compatible modalities with new styles for RGB-IR Re-ID. Besides, more new modalities perform better with stronger promotion to explore potential information.

Evaluation of SSC. We evaluate the designed image-level losses using experiments with the same option of modalities in the last three groups. Results demonstrate SSC compensates for the limitation of ISA we mentioned in Section 3.2 and benefits to exploit more related information in image space. However, another observation is the improvement seems slim with only the R2I or I2R modality. One reasonable analysis is there are fewer parameters (fewer ISA modules) to optimize when only one new modality is generated, which limits the effect of SSC.

Evaluation of SSL and SSL$^+$. In this paper, we propose two different image quantification metrics along with two losses separately for $\{X_t, X_a\}$ and $\{X_s, X_a\}$. To research how the two losses work with different pairs of images, we evaluate $\mathcal{L}_{SSL}$ and $\mathcal{L}_{SSL^+}$ in detail as shown in Table 5. Actually, performances of different options are almost the same on the RegDB dataset but vary on the larger and more challenging SYSU-MM01 dataset. Results show that both SSL

Index	X_s–X_a	X_t–X_a	r = 1	mAP
1	$\mathcal{L}_{SSL}$	$\mathcal{L}_{SSL}$	58.88	56.04
2	$\mathcal{L}_{SSL^+}$	$\mathcal{L}_{SSL^+}$	59.98	57.25
3	$\mathcal{L}_{SSL^+}$	×	55.34	52.22
4	×	$\mathcal{L}_{SSL}$	54.75	52.46
5 (MSA)	$\mathcal{L}_{SSL^+}$	$\mathcal{L}_{SSL}$	**63.13**	**59.22**

Table 5: Evaluation of $\mathcal{L}_{SSL}$ and $\mathcal{L}_{SSL^+}$ on SYSU-MM01 dataset under *all-search* mode.

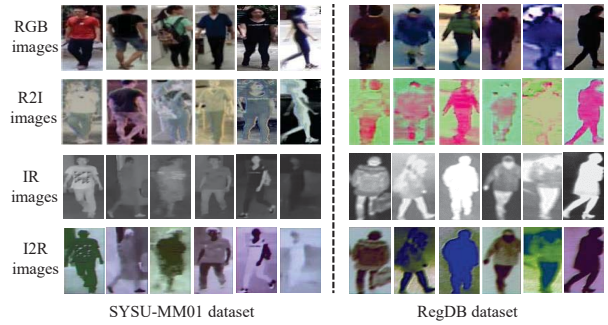

Figure 5: Visualization of X_{r2i} and X_{i2r} images on two datasets.

and SSL$^+$ are valid to guide the style adaptation though SSL$^+$ provides a more powerful constraint. Moreover, SSL fits $\{X_t, X_a\}$ better due to the content differences between them, which weaken the pixel-based SSL$^+$ designed in a global view as described in Section 3.2. In a word, the patch-view SSL fits images with large content difference and the SSL$^+$ is more suitable for images with the same content information to strengthen the constraint from a global view.

4.4 Discussions

Image-level visualization. In this part, we show images of the genearted R2I and I2R modalities. As shown in Fig. 5, they inherit content information from source images, meanwhile the luminance, contrast as well as the 'tone' of color are changed to a new form. Specially, the I2R images are more colorful than the source IR images, profiting from the knowledge siphoned off from RGB images by the designed SSC module. Moreover, although they look not perfect enough, the improvements in Section 4.2 support our initial motivation: the high-similarity fake images are not necessary.

Feature-level visualization. One key point of MSA is we achieve the style adaptation throughout in image space. To explore how the image-level operations affect the feature space after a deep CNN backbone, we visualize the feature space using t-SNE [Maaten and Hinton, 2008] as shown in Fig. 6. Compared with the baseline, MSA introduces R2I and I2R, and features from initial RGB and IR modalities are clustered mainly based on identity. In Fig. 6b, the generated X_a and its corresponding X_s are almost overlapped because of the simple structure of ISA. In Fig. 6c, X_a can be distinguished from X_s as shown in the dashed boxes and can extract more potential information about X_s. Accordingly, the feature space is redistributed due to the addition of new modalities, during which more potential connections and

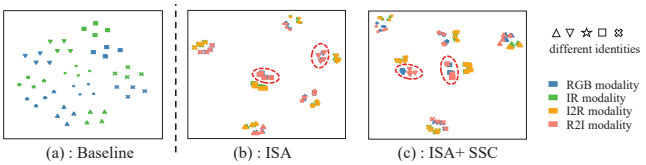

Figure 6: Visualization of feature space. (a) Results of the baseline model with two initial modalities. (b) Results of model trained with ISA. (c) Results of model trained with ISA and SSC. With SSC, X_a seems more independent from X_s. (Please view in color.)

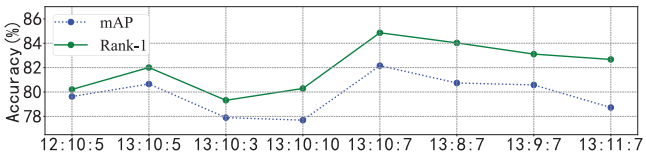

Figure 7: Performances of MSA with respect to different proportions of α: β: γ on RegDB dataset.

differences between RGB and IR modalities are explored because of the similarity among the four modalities. As a result, the variations between RGB and IR modalities are effectively reduced in feature space.

Parameters analysis. During training, three key parameters α, β and γ in $\mathcal{L}_{total}$ will guide the optimization of the whole network. Fig. 7 shows the performances of MSA when the three parameters are set as different values. Experiments show that MSA performs optimally when the proportions of them are 13:10:7 on the RegDB dataset.

Model analysis. MSA aims to achieve a highly compact framework with no need for high-quality generated images. To this end, we evaluate the processing FLOPs and model parameters of MSA. For a 288 × 144 image, the FLOPs of MSA are 5.1689G, only 0.0072G more than the baseline. Moreover, the extra model parameters are less than 0.0001M with a total of 24.32M. Additionally, the model parameters of the backbone of cm-SSFT [Lu *et al.*, 2020] (SOTA) are more than 70M. As a result, MSA achieves 20.30% and 17.25% improvements compared with the baseline concerning rank-1 and mAP on the SYSU-MM01 dataset.

5 Conclusion

A modality-aware style adaptation (MSA) framework is proposed for the RGB-IR Re-ID, which introduces two new modalities to assist the exploration of modality-invariant features. We design a highly compact framework, containing a trainable feature-free image reconstruction structure to generate images, two image-based style quantification metrics and two image-level style distance losses to assist optimization of the whole network. As a result, the generated images significantly boost the performance of RGB-IR Re-ID.

Acknowledgments

This work is supported by National Key R&D Program of China (No.2020AAA0108904), Science and Technology Plan of Shenzhen (No.JCYJ20190808182209321).

References

[Choi *et al.*, 2020] 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*, pages 10257–10266, 2020.

[Dai *et al.*, 2018] Pingyang Dai, Rongrong Ji, Haibin Wang, Qiong Wu, and Yuyu Huang. Cross-modality person re-identification with generative adversarial training. In *IJCAI*, pages 677–683, 2018.

[Feng *et al.*, 2019] Zhanxiang Feng, Jianhuang Lai, and Xiaohua Xie. Learning modality-specific representations for visible-infrared person re-identification. *IEEE Transactions on Image Processing*, 29:579–590, 2019.

[Hao *et al.*, 2019] Yi Hao, Nannan Wang, Jie Li, and Xinbo Gao. Hsme: hypersphere manifold embedding for visible thermal person re-identification. In *AAAI*, pages 8385–8392, 2019.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[Jia *et al.*, 2020] Mengxi Jia, Yunpeng Zhai, Shijian Lu, Siwei Ma, and Jian Zhang. A similarity inference metric for rgb-infrared cross-modality person re-identification. *arXiv preprint arXiv:2007.01504*, 2020.

[Krizhevsky *et al.*, 2012] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, pages 1097–1105, 2012.

[Li *et al.*, 2020] Diangang Li, Xing Wei, Xiaopeng Hong, and Yihong Gong. Infrared-visible cross-modal person re-identification with an x modality. In *AAAI*, pages 4610–4617, 2020.

[Liu *et al.*, 2018] Hong Liu, Wei Shi, Weipeng Huang, and Qiao Guan. A discriminatively learned feature embedding based on multi-loss fusion for person search. In *ICASSP*, pages 1668–1672, 2018.

[Liu *et al.*, 2020] Hong Liu, Zhisheng Lu, Wei Shi, and Juanhui Tu. A fast and accurate super-resolution network using progressive residual learning. In *ICASSP*, pages 1818–1822, 2020.

[Lu *et al.*, 2020] 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*, pages 13379–13389, 2020.

[Maaten and Hinton, 2008] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9:2579–2605, 2008.

[Nguyen *et al.*, 2017] 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.

[Wang *et al.*, 2004] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[Wang *et al.*, 2019a] Guanan 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*, pages 3623–3632, 2019.

[Wang *et al.*, 2019b] Zhixiang Wang, Zheng Wang, Yinqiang Zheng, Yungyu Chuang, and Shinichi Satoh. Learning to reduce dual-level discrepancy for infrared-visible person re-identification. In *CVPR*, pages 618–626, 2019.

[Wang *et al.*, 2020] Guanan Wang, Tianzhu Zhang, Yang Yang, Jian Cheng, Jianlong Chang, Xu Liang, and Zengguang Hou. Cross-modality paired-images generation for rgb-infrared person re-identification. *arXiv preprint arXiv:2002.04114*, 2020.

[Wu *et al.*, 2017] Ancong Wu, Weishi Zheng, Hongxing Yu, Shaogang Gong, and Jianhuang Lai. Rgb-infrared cross-modality person re-identification. In *ICCV*, pages 5380–5389, 2017.

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

[Ye *et al.*, 2019a] Mang Ye, Xiangyuan Lan, and Qingming Leng. Modality-aware collaborative learning for visible thermal person re-identification. In *ACMMM*, pages 347–355, 2019.

[Ye *et al.*, 2019b] Mang Ye, Xiangyuan Lan, Zheng Wang, and Pong C Yuen. Bi-directional center-constrained top-ranking for visible thermal person re-identification. *IEEE Transactions on Information Forensics and Security*, 15:407–419, 2019.

[Ye *et al.*, 2020a] Hanrong Ye, Hong Liu, Fanyang Meng, and Xia Li. Bi-directional exponential angular triplet loss for rgb-infrared person re-identification. *arXiv preprint arXiv:2006.00878*, 2020.

[Ye *et al.*, 2020b] 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*, pages 229–246, 2020.

[Ye *et al.*, 2020c] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *arXiv preprint arXiv:2001.04193*, 2020.

[Zhang *et al.*, 2019] Zhifei Zhang, Zhaowen Wang, Zhe Lin, and Hairong Qi. Image super-resolution by neural texture transfer. In *CVPR*, pages 7982–7991, 2019.