IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 30, 2020, accepted January 10, 2021, date of publication January 14, 2021, date of current version January 22, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3051637

AS-RIG: Adaptive Selection of Reconstructed Input by Generator or Interpolation for Person Re-Identification in Cross-Modality Visible and Thermal Images

JIN KYU KANG, MIN BEOM LEE, HYO SIK YOON, AND KANG RYOUNG PARK, (Member, IEEE)
Division of Electronics and Electrical Engineering, Dongguk University, Seoul 04620, South Korea

Corresponding author: Kang Ryoung Park (parkgr@dongguk.edu)

This work was supported in part by the Ministry of Science and ICT (MSIT), Korea, under the Information Technology Research Center (ITRC) Support Program supervised by the Institute for Information and Communications Technology Promotion (IITP) under Grant IITP-2020-2020-0-01789, and in part by the National Research Foundation of Korea (NRF) funded by the MSIT through the Basic Science Research Program under Grant NRF-2020R1A2C1006179 and Grant NRF-2019R1A2C1083813.

ABSTRACT Multimodal camera-based person re-identification (ReID) is important in the field of intelligent surveillance. Thermal cameras can solve the problem in that visible-light cameras cannot acquire the valid feature information of a person under poor illumination conditions. However, thermal cameras usually have lower frame resolution than visible-light cameras. To overcome this problem, we propose an adaptive selection of reconstructed input by generator or interpolation (AS-RIG) method, which can adaptively select the generative adversarial network (GAN), or an interpolation method (bi-linear or bi-cubic). AS-RIG automatically selects a resolution-model using the mean-squared error (MSE), feature distance (FD), and structural similarity (SSIM). To verify the performance of our proposed method, two open databases are used: the DBPerson-Recog-DB1 and Sun Yat-set University multiple modality Re-ID (SYSU-MM01). Infrared frames from both databases are resized to be smaller than the original ones for experimentation. Experimental results show that our generator outperforms traditional interpolation methods. In addition, the person ReID experimental results demonstrate that AS-RIG outperforms non-adaptive selection methods and state-of-the-art methods.

INDEX TERMS Person Re-ID, convolutional neural network (CNN), super-resolution (SR), GAN.

I. INTRODUCTION

Person re-identification (ReID) aims to match a specific person having varying viewpoints and poses from two or more frames that are captured from more than one camera. Compared to the tracking algorithm, this is difficult because it is an environment that has non-continuity with respect to the axis in time. Recently, research on person ReID has been performed owing to the need for intelligent surveillance systems [1]. Research to re-identify a specific person using different visible-light cameras (CCTV) in the daytime environment is the main focus [2]–[4]. However, visible-light cameras are vulnerable to illumination conditions. In the case of intelligent surveillance systems, there is a functional meaning at night time, when the crime incidence rate is higher than in the day time. In order to solve this problem, there has been much interest in research on person ReID using a multimodal camera that combines a visible-light camera and an infrared camera [5]–[8]. Multimodal cameras are free from illumination conditions, but it is relatively difficult to utilize the input data of different properties. Compared to relatively high-resolution visible-light cameras, thermal cameras have a lower resolution, and the person region captured by a thermal camera in an intelligent surveillance system is smaller than that by the visible-light camera, which causes the degradation of quality of person region in the thermal camera image. In general, low-resolution thermal image can be resized using an interpolation method such as

The associate editor coordinating the review of this manuscript and approving it for publication was Tao Zhou.

bi-cubic or bi-linear, but it has the limitation of performance enhancement, and this is the major challenge of our research. We aim to improve the result of person ReID compared to the interpolation method by utilizing the generator model. The superiority of adaptive selection of reconstructed input by generator or interpolation (AS-RIG) proposed in our paper is demonstrated using two types of open databases (DBPerson-Recog-DB1 and SYSU-MM01). This means that AS-RIG has successfully improved the physical properties of the thermal camera using the generator. In addition, by performing experiments, we found that the generator model is not always superior to the interpolation method. Following this discovery, we devised a system that can be applied to the cross-modality person ReID algorithm by evaluating the reconstructed thermal frames using the mean-squared error (MSE), feature distance (FD), and structural similarity (SSIM). The main contributions of this paper are as follows:

- We propose an AS-RIG method, which can improve the accuracies of person ReID by adaptively selecting a generator or bi-cubic to reconstruct the input data.
- By performing experiments, we discovered that the generative model is not always superior to the interpolation methods for person ReID.
- AS-RIG improves the performance of person Re-ID by adaptively selecting the reconstructed thermal frame or interpolated thermal frame using the MSE, FD, and SSIM.
- For ease of comparison, we made the proposed algorithm with models available by other research [41].

The remainder of this paper is as follows. In Section II, we analyze the previous research on person ReID by dividing it into three parts (person ReID using a visible-light camera, a multimodal camera, and generative model). In Section III, we explain in detail our propose method. Then, in Section IV, we present the experimental results with analyses, and we present the conclusions in Section V.

II. RELATED WORK

As a basic study on person ReID, Huang and Russell [11] proposed the Bayesian method for tracking the same object from various camera perspectives. Subsequently, many studies on person ReID based on a visible-light camera were introduced [12]–[18], [45], [48]–[50]. In addition, research based on multimodal cameras was also conducted to overcome the limitations of visible-light cameras and to improve accuracy [19]–[23]. In recent studies, person ReID has been interpreted from various perspectives based on generators [25], [27]–[30].

A. PERSON ReID USING VISIBLE-LIGHT CAMERA

Person ReID on a visible-light camera focuses on solving the problem of varying poses and viewpoints. Farenzena *et al.* [12] designed a three-phase process for a person ReID. This is one of the early approaches to finding the axes of symmetry and asymmetry parts in the frame of a person, and feature matching by extracting the features of pedestrians from each part. Subsequently, deep learning has been applied to person ReID, and various studies have been reported [5], [6]. Ahmed *et al.* [13] proposed a method to calculate the probability of similarity with softmax obtained from the concatenated feature map, which is extracted through convolutional layers from the input frame. Unlike the deep models for person ReID [14], [15], which proposed the verification loss between positive and negative frames for an anchor, Cheng *et al.* [16] proposed a person ReID that is based on a triplet loss function that can reduce intra-class variation. Hermans *et al.* [17] employed a batch hard to improve the previous triplet loss. The batch hard consists of a hardest negative frame ranked higher and a hardest positive frame ranked lower for the anchor frame. Chen *et al.* [18] proposed a quadruplet loss comprising four batches by adding an extra negative frame. It was reported that the quadruplet loss is more effective than triplet loss in increasing inter-class variation and reducing intra-class variation.

Person ReID using the generative model proposed new perspectives of person ReID by generating or transferring the primary issues such as varying pose, viewpoint, illumination, and data type. Zhong *et al.* [25] proposed CamStyle as a method to solve one of the person ReID challenges and frame style variations by different cameras. CamStyle is based on a cycle-consistent generative adversarial network (CycleGAN) [26], and it is introduced to solve the differences between camera styles by providing data augmentation. Wei *et al.* [27] proposed a person transfer generative adversarial network (PTGAN) that transfers the background and mean of the lighting in order to focus on the identity of a person. Ge *et al.* [28] reported that the pose variation of frames is a key challenge for learning robust person features. To solve this challenge, we proposed a feature distilling generative adversarial network (FD-GAN) that transfers two input frames to a target pose. Qin *et al.* [45] proposed the pedestrian ReID based on super-resolution images (SRPRID) which consisted of the super-resolution sub-network and ReID sub-network. In [48], authors propose the method of multi-scale learning for low resolution person re-identification, and Wang *et al.* proposed the method of scale-adaptive low-resolution person re-identification based on learning a discriminating surface [49]. In addition, Jing *et al.* propose the method of super-resolution person re-identification with semi-coupled low-rank discriminant dictionary learning [50]. However, all these researches did not deal with the cross-modality person ReID issue, but our research deals with the cross-modality person ReID issue between visible and thermal cameras.

B. PERSON ReID USING MULTIMODAL CAMERA

Person ReID using a multimodal camera aims to solve extremely illumination condition problems as well as varying poses and viewpoints. Wu *et al.* [19] proposed a person ReID that combines a visible-light camera and a depth camera.

In this study, invariant body shape and skeleton information freely under extremely illumination conditions and color change were applied to person ReID using a depth camera. There are also studies on person ReID using multimodal cameras that combine visible-light and infrared cameras [5]–[8], [21]. Ye *et al.* [20] proposed an end-to-end network by applying the identity loss and ranking loss based on the feature extraction of two different types of frames using each stream network. Kang *et al.* [22] proposed a one-stream network for person ReID using a single input data model. The single input data model is three-dimensional (3D) data composed of inter-channel pairs and intra-channel pairs based on visible-light frames and infrared frames. Lin and Li [23] introduced the pentaplet loss by applying the triplet loss introduced in [24] to multimodal data conditions. Pentaplet loss is a function that is designed so that the cross-modality variation and intra-modality variation are simultaneously trained by dividing them into five subsets: anchor (visible-light frames), negative (inter-class of visible-light frames), cross-modality negative (intra-class of visible-light frames), positive (inter-class of infrared frames), and cross-modality positive (intra-class of infrared frames).

Person ReID on a multimodal camera also used a generative model to solve the discrepancies related to heterogeneous data. Zhang *et al.* [29] introduced a study on person ReID under poor illumination conditions. This research proposed a Teacher-Student GAN (TS-GAN) that is capable of transferring RGB to infrared data in order to resolve the cross-modality discrepancy of features from frames captured by an infrared camera on night and a visible-light camera during the daytime. Choi *et al.* [30] aimed to reduce cross-modality discrepancies. Hierarchical cross-modality disentanglement (HI-CMD), consisting of identity preserving person image generation (ID-PIG) and hierarchical feature learning (HFL), reported that it can simultaneously reduce cross-modality and intra-modality discrepancies using ID-discriminative factors and ID-excluded factors. In [51], Ye *et al.* proposed a novel dynamic dual-attentive aggregation (DDAG) learning method by mining both cross-modality graph-level contextual cues and intra-modality part-level for visible-infrared (VI)-ReID. In [8], they proposed a homogeneous augmented tri-modal (HAT) learning method for VI-ReID. In their method, an auxiliary grayscale modality is generated from homogeneous visible images without additional training process. In [52], authors newly proposed a modality-aware collaborative ensemble (MACE) learning method with middle-level sharable two-stream network (MSTN) for VT-ReID. It handles the modality-discrepancy in both feature level and classifier level.

In [53], Ye *et al.* proposed a dual-path network with a new bi-directional dual-constrained top-ranking (BDTR) loss to learn discriminative feature representations. Although it is not about the research of person ReID, authors newly proposed an instance-wise softmax embedding, which directly executes the optimization over the augmented instance features with the binary discrimination softmax encoding [54]. Although they are not about the researches of person ReID, Luo *et al.* proposed a new dimensionality reduction (DR) method, termed local geometric structure Fisher analysis (LGSFA), for HSI classification of hyperspectral imagery [43]. In addition, Shi *et al.* proposed a novel unsupervised DR method called local neighborhood structure preserving embedding (LNSPE) for HSI classification of hyperspectral imagery [44].

Table 1 summarizes the advantages and disadvantages of the methods proposed in the conventional studies and this study for person ReID.

III. PROPOSED METHOD

In general, the image resolution of pedestrian region detected by a visible-light camera is higher than that by a thermal camera. Owing to this problem, the reconstruction of data size for the thermal frame is essential for the cross-modality person ReID. Based on this statement, we make the following assumptions: (1) we expect that if the result of the super-resolution (SR) generator is better than the bi-cubic, which is one of the traditional interpolation methods, the person ReID performance can be improved. (2) If the SR generator is not always better than the bi-cubic, we can further improve the performance of person ReID using the adaptive selection method. Below, we introduce the proof of this assumption. First, in Section III.*A*, we briefly review the overall AS-RIG procedure. We describe the reconstructing input data by generator in Section III.*B*. The criteria for adaptive selection and the baseline person ReID model are described in Sections III.*C* and III.*D*, respectively.

A. OVERALL PROCEDURE OF AS-RIG

Fig. 1 shows the overall procedure of the AS-RIG method proposed in this paper. AS-RIG aims to improve the person ReID between visible-light frames captured during daytime and thermal frames captured at night. In general, cross-modality person ReID uses visible-light frames and thermal frames as input data. In this case, the key goal of AS-RIG is to reconstruct the input size of the thermal data using the SR generator. Because person ReID is based on unpaired data, the generator of AS-RIG is composed of CycleGAN [26].

However, there are cases where the output of the generator loses its identify features. To overcome this problem, we designed a generator and bi-cubic to be selected according to each scenario. We used MSE, FD, and SSIM for adaptive selection to evaluate the reliability of the reconstructed thermal frame using a generator and resized the thermal frame using a bi-cubic. Then, IPVT-1 [22] is created using the thermal frame selected through the preprocessing and the visible-light frame, which is one of the anchor sets. IPVT-1, which is constructed as a result of adaptive selection, is more advantageous in extracting cross-modality features than IPVT-1, which is not. That is, the Person ReID network can be facilitated in the direction of decreasing intra-class variation and increasing inter-class variation.

TABLE 1. Comparison of proposed and previous research on person ReID.

Category		Advantage	Disadvantage
Visible-light camera	Multi-channel with improved triplet loss [16]	The conventional triplet loss method is improved by learning both all and part of the human body	They are susceptible to extremely illumination conditions
	Defense of the triplet loss [17]	To improve the triplet loss, a batch hard consisting of hardest negative and hardest positive was proposed	
	Quadruplet loss [18]	The intra-class distance is reduced and the inter-class distance is increased by adding a negative-negative pair in the conventional triplet loss	
	CamStyle [25]	CamStyle is introduced to solve disparities of camera style by providing data augmentation	
	PTGAN [27]	PTGAN is designed to transfer the background and mean of lighting in order to focus on the identity of person	
	FD-GAN [28]	FD-GAN is designed to solve the pose variation of frames	
	SRPRID [45]	SRPRID is designed to solve the low-resolution of frames	
Multimodal camera	Visible-depth frames [19]	Using depth information to overcome the extreme lighting and change of clothes	They do not consider the low-resolution frame using a thermal camera
	Dual-constrained top-ranking [20]	To learn discriminative cross-modality features, a dual path network is designed using the dual-constrained top-ranking loss	
	IPVT-1 [22]	IPVT-1 is introduced for training with one stream network to extract features from heterogeneous data	
	HPILN [23]	HPILN is designed with a pentaplet loss to train cross-modality variation and intra-modality variation simultaneously	
	TS-GAN [29]	TS-GAN can resolve the cross-modality discrepancy of features from frames by transferring RGB to infrared data	
	HI-CMD [30]	HI-CMD can simultaneously reduce cross-modality and intra-modality discrepancies	
	AS-RIG (proposed method)	- Consider low-resolution frames using a thermal camera - Improve the person ReID field by adaptively selecting a generator or bi-cubic to resize the input data	Additional training is necessary for a generative adversarial network

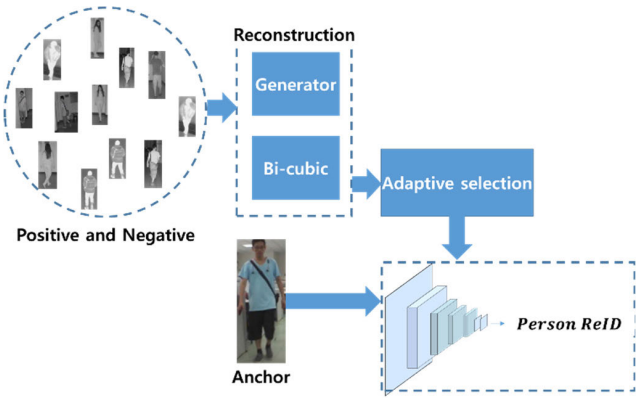


FIGURE 1. Overall procedure of AS-RIG.

B. RECONSTRUCTING INPUT DATA USING A GENERATOR

Thermal frames require reconstruction before the cross-modality person ReID. This is because the resolution of thermal frames is smaller than that of visible-light frames. To solve this problem, we use the CycleGAN [26] that is suitable for the unpaired dataset.

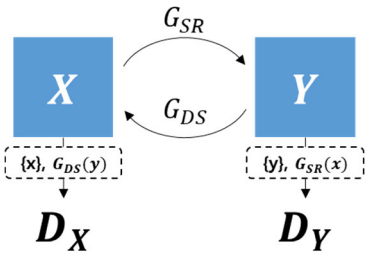


FIGURE 2. Structure of generator in CycleGAN.

Fig. 2 shows the structure of CycleGAN for reconstructing thermal frames. Our goal is to train the mapping function between two domains, X and Y, given downsized thermal frames $\{x_i\}_{i=1}^{N}$ where $x_i \in X$ and original thermal frames $\{y_j\}_{j=1}^{M}$ where $y_j \in Y$. Our generator includes two mapping functions $G_{SR} : X \rightarrow Y$ and $G_{DS} : Y \rightarrow X$. G_{DS} is only used to train G_{SR}, and G_{SR} is needed for our AS-RIG. To train mapping functions G_{SR} and G_{DS}, we used two adversarial discriminators D_X and D_Y, where D_X is designed to differentiate between {x} and the generated frames G_{DS}(Y)}, and

D_Y is designed to differentiate between {y} and the generated frames $\{G_{SR}(X)\}$. In addition, we used the loss function to CycleGAN as shown in Equation (1).

$$\mathcal{L}(G_{SR}, G_{DS}, D_X, D_Y) = \mathcal{L}_{GAN}(G_{SR}, D_Y, X, Y) + \mathcal{L}_{GAN}(G_{DS}, D_X, Y, X) + \lambda \mathcal{L}_{cyc}(G_{SR}, G_{DS}) \quad (1)$$

$\mathcal{L}_{GAN}$ is the adversarial loss to match the distribution of the generated frames with the distribution of data in the target domain. One of $\mathcal{L}_{GAN}$, $\mathcal{L}_{GAN}(G_{SR}, D_Y, X, Y)$, can be defined as

$$\mathcal{L}_{GAN}(G_{SR}, D_Y, X, Y) = \mathbb{E}_{y\sim p_{data}(y)}[logD_Y(y)] + \mathbb{E}_{x\sim p_{data}(x)}[log(1 - D_Y(G_{SR}(x)))] \quad (2)$$

where G_{SR} learns so that $G_{SR}(x)$ maintains its identity, but has a similar high-resolution quality as domain Y. At this time, D_Y attempts to distinguish between generated frames $G_{SR}(x)$ and real samples, y. That is, $D_Y(G_{SR}(x)$ attempts to converge to 0 and $D_Y(y)$ try to converge to 1. From that, adversarial (discriminator) loss ($\mathcal{L}_{GAN}(G_{SR}, D_Y, X, Y)$) can converge to 0 when the discriminator sufficiently trains, which does not mean that the model is collapsing.

$\mathcal{L}_{cyc}$ is the cycle consistency loss that prevents the trained mapping function G_{SR} to contradict G_{DS}, and we express the equation as:

$$\mathcal{L}_{cyc}(G_{SR}, G_{DS}) = \mathbb{E}_{x\sim p_{data}(x)}\left[\|G_{DS}(G_{SR}) - x\|_1\right] + \mathbb{E}_{y\sim p_{data}(y)}[\|G_{SR}(G_{DS}) - y\|_1] \quad (3)$$

Equation (3) closely matches the reconstructed frames $G_{DS}(G_{SR})$ and $G_{SR}(G_{DS})$ to the input frames x and y, respectively.

C. CRITERIA FOR ADAPTIVE SELECTION

As described in Section III.*A*, there are cases where the generated thermal frames, which are from CycleGAN, lose their identify features. To overcome this problem, we propose an adaptive selection that chooses a thermal frame, whether generated by CycleGAN or resized by bi-cubic, using the following methods.

1) CASE-I: MEAN-SQUARED ERROR

The MSE is a mathematical function that is used to estimate mapping data against baseline data. In our case, MSE is the average of the difference value of each pixel between the generated thermal frame and the resized thermal frame. In other words, it measures the visible similarity of their input data. MSE is expressed as follows:

$$\text{MSE} = \frac{1}{WH}\sum_{i=0}^{H-1}\sum_{j=0}^{W-1}[G_{SR}(x(i,j)) - I_{bi-cubic}(i,j)]^2 \quad (4)$$

In Equation (4), $G_{SR}(x)$ is a generated thermal frame and $I_{bi-cubic}$ is a resized frame obtained using a bi-cubic, which is one of the interpolation methods. H and W denote the height and width of the frame, respectively.

2) CASE-II: FEAUTRE DISTANCE

The FD is obtained by calculating the Euclidean distance between the output nodes of the last convolutional layer in the deep network using the generated thermal frame and resized thermal frame. The output nodes of the last convolutional layer are high-dimensional feature sets that include the identity of each person.

The Euclidean distance of this feature set means that the generated thermal frame is similar to the original frame. We define the FD equation as

$$\text{FD} = \sqrt{\sum_{k=1}^{n}[O_{last\_conv}(G_{SR}(x)) - O_{last\_conv}(I_{bi-cubic})]^2} \quad (5)$$

In Equation (5), O_{last_conv} denotes the output nodes of the last convolutional layer in the deep network, and n is the number of output nodes.

3) CASE-III: STRUTURAL SIMILARITY

The SSIM [31] is used to measure the similarity between two frames. This method compares local patterns of pixel intensities that have been reconstructed for luminance and contrast. Compared to other methods, such as MSE or PSNR [32], the difference is that it estimates the absolute error. The SSIM is expressed as follows:

$$\text{SSIM} = \frac{(2\mu_G\mu_I + C_1)(2\sigma_{GI} + C_2)}{(\mu_G^2 + \mu_I^2 + C_1)(\sigma_G^2 + \sigma_I^2 + C_2)} \quad (6)$$

Equation (6) combines the three comparisons of luminance comparison, constant comparison, and structure comparison for two signals $G_{SR}(x)$ and $I_{bi-cubic}$. In Equation (6), μ_G and σ_G^2 denote the average and variance of the generated thermal frame $G_{SR}(x)$, respectively. μ_I and σ_I^2 denote the average and variance of the resized thermal frame $I_{bi-cubic}$, respectively. σ_{GI} denotes the covariance of $G_{SR}(x)$ and $I_{bi-cubic}$. $C_1 = (k_1L)^2$ and $C_2 = (k_2L)^2$ are two variables that are used to stabilize the division with a weak denominator. L denotes the dynamic range of pixel values. $k \ll 1$ is a small constant.

D. BASELINE PERSON ReID MODEL

To prove that our proposed AS-RIG can improve the performance of the cross-modality person ReID, we choose a baseline network, which consists of a one-stream network with IPVT-1 [22]. This is because the network was shown to have a higher performance of cross-modality person ReID than others on two open-databases. That is, Ref. [22] is appropriate to show that the performance can be improved using AS-RIG.

Fig. 3 displays the structure of the cross-modality person ReID [22]. Because this network is a one-stream network, we used IPVT-1 composed of inter-channel pairs for heterogeneous data as a single input. ResNet-50 [33] was used for the convolutional neural network structure.

As indicated in Fig. 3, the number of outputs of the fully connected layer in the network defines two nodes. This is

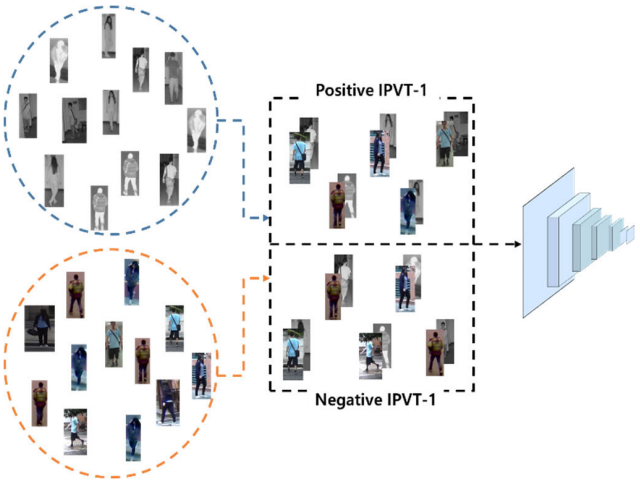


FIGURE 3. Structure of cross-modality person ReID.

to calculate the probability of a similarity between identities using heterogeneous data as one input.

IV. EXPERIMENTAL RESULTS

To demonstrate the performance of AS-RIG, we experimented with two widely used open databases, DBPerson-Recog-DB1 [9] and Sun Yat-set University multiple modality Re-ID (SYSU-MM01) [10]. There are two main types of experiments. First, we evaluated the quality of the frame generated by the cycleGAN-based SR generator described in Section III.*B*. Second, we compared the performance of person ReID using AS-RIG with other methods.

A. EXPERIMENTAL DATABASES

1) DBPERSON-RECOG-DB1

As indicated in Table 2, DBPerson-Recog-DB1 is a database that is acquired by a multimodal camera composed of a visible-light camera (Logitech C600 [34]) and a thermal camera (FLIR Tau2 [35]). This database consists of *8240* frames of visible-light frames and thermal frames, each including various camera views such as front, side, and back views of 412 people. It is a database that is acquired outdoors, and the average size of visible-light frames is $37 \times 102 \times 3$ pixels, and the average size of the thermal frames is $42 \times 112 \times 3$ pixels. Fig. 4 shows sample frames of DBPerson-Recog-DB1 used to verify the AS-RIG. As shown in the sample frames, DBPerson-Recog-DB1 is a database that is acquired from the same viewpoint using a visible-light camera and a thermal camera. To verify the AS-RIG performance, we used a two-fold cross-validation method by dividing it into two subsets, as shown in Table 2. However, this dataset has an average size of visible-light frames similar to that of thermal frames because a low-resolution visible-light camera is used. Therefore, in order to fit the experimental environment that we designed, the size of the thermal frames was arbitrarily downsized to one quarter of its original size.

TABLE 2. Description of DBPerson-Recog-DB1.

	DBPerson-Recog-DB1	
	Subset 1	Subset 2
# of identities	206	206
# of frames	4120	4120
Environments	Outdoors, varying viewpoint, multimodal camera (visible-light & thermal camera), paired data	

(a) (b)

FIGURE 4. Sample frames of DBPerson-Recog-DB1.

2) SYSU-MM01

As indicated in Table 3, SYSU-MM01 is also a cross-modality dataset acquired by the Kinect V1 and IR cameras. SYSU-MM01 consists of a total of 44,745, frames including 30,071 visible-light frames and 15,792 infrared frames. This database has various viewpoints such as front, side, and back views that are similar to DBPerson-Recog-DB1. The average size of the visible-light frames is $112 \times 284 \times 3$ pixels, and the average size of the infrared frames is $108 \times 303 \times 3$ pixels. However, unlike DBPerson-Recog-DB1, it has several characteristics. First, it is a database that is composed of a visible-light camera and an infrared camera independently. Fig. 5 shows a sample frame of SYSU-MM01. As shown in Figs. 5(a) and (b), unlike Fig. 4, it can be seen that the visible-light frames and infrared frames are unpaired sets. Second, it was captured both outdoors and indoors. Because of the first and second characteristics described above, SYSU-MM01 makes person ReID more difficult because there are more variables than DBPerson-Recog-DB1. Third, the databases provided by [10] are already classified as training, set, validation set, and test set. Therefore, we used this configuration as they were to determine the performance of AS-RIG in the verification experiment. Fourth, because this dataset used an infrared camera rather than a thermal camera, the average size of infrared frames is similar to that of visible-light frames. Therefore, we downsized infrared frames to one eighth of the size to fit the designed experimental environment.

B. TRAINING

To train the generator described in Section III. *B*, CycleGAN was scratch trained using the Pytorch framework (version 1.2 [42]). For generator training, we used adaptive moment estimation [36] as the optimization function, which combines the advantages of two methods: AdaGrad [37] to deal with sparse gradients and the ability of RMSProp [38] to deal with non-stationary objectives. All of the experiments were

TABLE 3. Description of SYSU-MM01.

	SYSU-MM01		
	Training set	Validation set	Test set
# of Identity	296	99	96
# of frames	30 213	3954	10 578
Environments	Outdoors and indoors, varying viewpoint, multimodal camera (visible-light & infrared camera), unpaired data		

(a) (b)

FIGURE 5. Sample frames of SYSU-MM01.

performed on a desktop computer having an Intel® Core™ i7-7700K CPU @ 3.60 GHz processor (4 cores), a main memory of 32 GB, and an NVIDIA GeForce GTX 1070 (1920 compute unified device architecture (CUDA) cores) with a graphics card that has a memory of 8 GB [39].

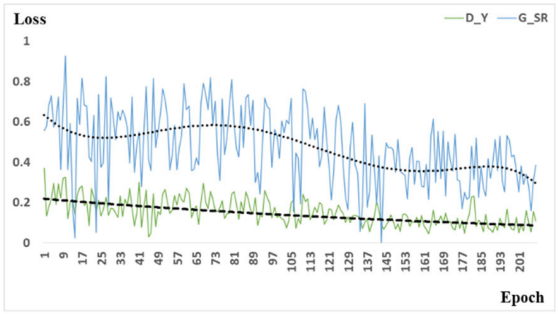
Fig. 6 shows the graphs of discriminator and generator losses according to training epochs. As shown in the graphs, it can be confirmed that the SR generator and discriminator were sufficiently trained. DBPerson-Recog-DB1 is to reconstruct from the average size of $11\times28\times3$ pixels to $42\times112\times3$ pixels, and SYSU-MM01 is to reconstruct from the average size $14 \times 38 \times 3$ pixels to $108 \times 303 \times 3$ pixels.

Both databases have small amounts of data for reconstructing high-resolution frames, but there are different data types. As indicated in Fig. 6, SYSU-MM01 converges faster than DBPerson-Recog-DB1. This means that it is more difficult to reconstruct high-resolution thermal frames than infrared frames. Fig. 7 shows the sample of bi-cubic and reconstructed thermal frames. Figs. 7 (a) and (b) are the results of DBPerson-Recog-DB1, and Figs. 7 (c) and (d) are the results of SYSU-MM01. For frames resized by bi-cubic, high-frequency elements were reduced, and there was a blur effect throughout the frames. However, in frames that are reconstructed by generator, the high-frequency elements were generated such that they are similar to the original frame, and the edges were clearly displayed.

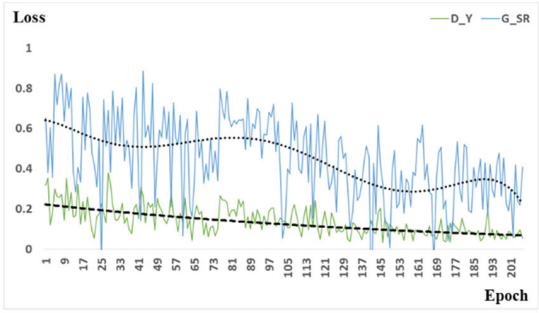
C. TESTING OF AS-RIG

1) DBPERSON-RECOG-DB1 (ABLATION STUDIES)

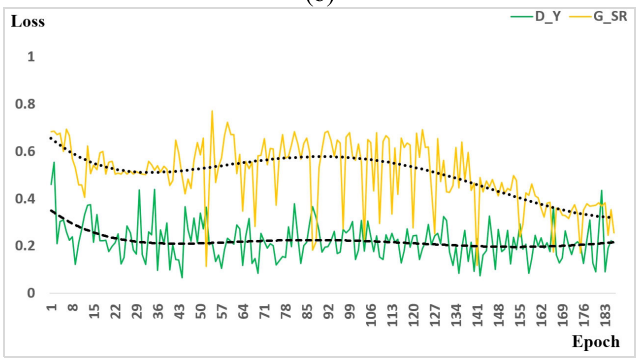
We conducted 12 experiments using DBPerson-Recog-DB1. First, the three experiments are the evaluation of person ReID using the interpolation of thermal frames in bi-linear or bi-cubic, and using reconstruction with the generator proposed in Section III. *B*. In the second experiment, the three adaptive selection methods in Section III. *C*, i.e., MSE, FD, and SSIM, were compared. Each method of adaptive selection is



(a)



(b)



(c)

FIGURE 6. Graphs of discriminator (D_Y) and generator losses (G_SR) in the training procedure: (a) first fold with DBPerson-Recog-DB1, (b) second fold with DBPerson-Recog-DB1, and (c) SYSU-MM01.

needed as a criterion for selecting interpolation and generator. To determine this criterion, we calculated the average value μ_C of each method from the training set. Based on the average value $\{\mu_C\}$, we define the threshold values A, B, and C that satisfy $A < B = \{\mu_C\} < C$.

We used Rank 1, Rank 10, Rank 20, and the mean average Precision (mAP) for the experimental evaluations. Rank N is the concept of evaluating the correct matching accuracy for cases containing data of the positive class (true positive data) out of N matching candidates. The mAP is the mean of the average precision scores for each anchor [40]. The average precision method indicates an area of the precision-recall graph, evaluates the identification algorithm. The mAP is expressed as follows:

$$\text{mAP} = \frac{\sum_{q=1}^{Q} AveP(q)}{Q} \tag{7}$$

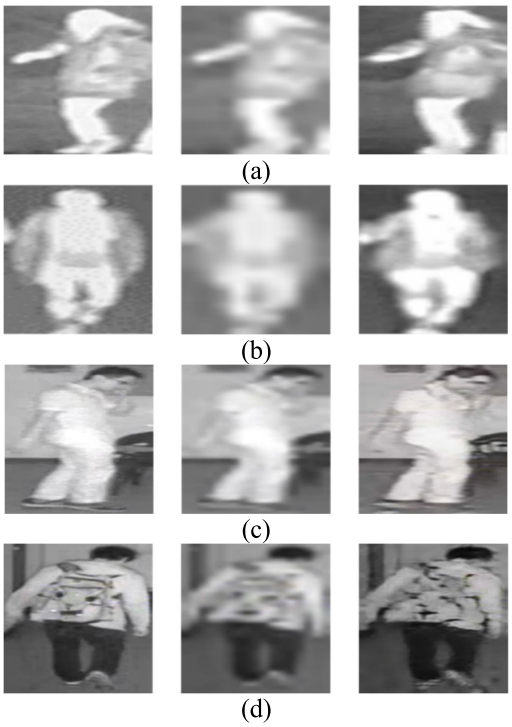


FIGURE 7. Sample of bi-cubic and reconstructed thermal frames: (a), (b) DBPerson-Recog-DB1 and (c), (d) SYSU-MM01. The left frame is the original frame, the middle frame is a result resized using the bi-cubic, and the right frame is a result reconstructed using the generator.

TABLE 4. Comparison of Person ReID according to the number of image pixels (unit: %).

Method		Rank 1	Rank 10	Rank 20	mAP
Original [22]		58.57	83.74	88.47	49.11
1/2 size of original image	Only bi-linear	56.12	87.72	83.84	40.61
	Only bi-cubic	57.82	88.91	94.25	41.62
	Only generator	52.55	85.49	91.77	37.79
	AS-RIG	**60.37**	**90.12**	**95.43**	**43.43**
1/4 size of original image	Only bi-linear	42.94	78.79	87.40	31.63
	Only bi-cubic	46.75	82.23	89.37	34.21
	Only generator	41.02	78.84	87.45	28.82
	AS-RIG	**57.60**	**88.96**	**94.13**	**41.34**
1/8 size of original image	Only bi-linear	35.36	60.43	76.87	24.29
	Only bi-cubic	36.98	61.79	78.22	24.78
	Only generator	38.87	60.97	79.29	26.12
	AS-RIG	**40.49**	**65.79**	**80.73**	**30.98**

In Equation (7), Q denotes the number of anchors and $AveP\,(q)$ indicates the average precision scores for each anchor.

In Table 4, we compared the accuracies of person ReID according to the number of image pixels. In this table, the original means the original size of image. As shown in this table, the accuracies are reduced according to the decrement of the number of image pixels which causes the information loss in image.

TABLE 5. Comparison of AS-RIG with DBPerson-Recog-DB1 (unit: %).

Method		DBPerson-Recog-DB1			
		Rank 1	Rank 10	Rank 20	mAP
Original [22]		58.57	83.74	88.47	49.11
Only bi-linear		42.94	78.79	87.40	31.63
Only bi-cubic		46.75	82.23	89.37	34.21
Only generator		41.02	78.84	87.45	28.82
AS-RIG: MSE	Threshold A	51.67	85.70	92.35	35.41
	Threshold B	51.14	85.56	92.16	34.96
	Threshold C	50.53	84.98	91.84	34.36
AS-RIG: SSIM	Threshold A	47.06	82.99	90.49	32.46
	Threshold B	54.27	86.60	93.13	37.91
	Threshold C	**57.60**	**88.96**	**94.13**	**41.34**
AS-RIG: FD	Threshold A	52.60	85.90	92.55	36.16
	Threshold B	51.31	85.32	91.99	34.91
	Threshold C	49.83	84.42	91.67	33.80

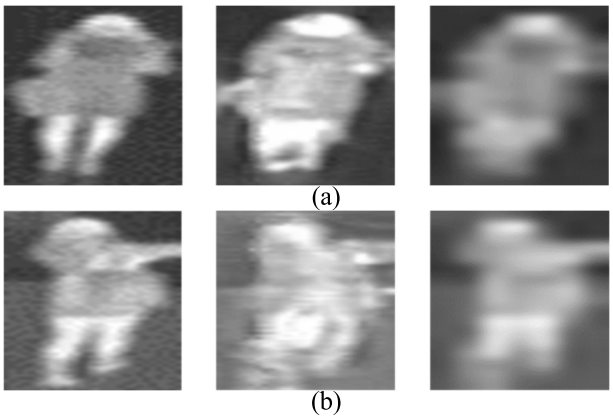


FIGURE 8. Examples of comparison bi-cubic with generator test result on DBPerson-Recog-DB1: (a) The left frame is the original frame, the middle frame is recorded rank 2 (reconstructed by generator), and the right frame is recorded rank 104 (resized by bi-cubic). (b) The left frame is the original frame, the middle frame is recorded rank 120 (reconstructed by generator), and the right frame is recorded rank 1 (resized by bi-cubic).

Nevertheless, our proposed method shows the higher accuracies than those by only bi-linear, bi-cubic, and generator of CycleGAN in all the cases of the reduction of image pixels. Even in the case of 1/2 size of original image, the rank 1, 10, and 20 by proposed AS-RIG are higher than those in the case of the original size of image. That is because the noises can be much reduced by the 1/2 size of original image compared to the original size of image, which causes the enhancement of accuracies.

In Table 5, we compared the accuracies according to various methods and thresholds in proposed AS-RIG, and also compared them with original image, and only bi-linear, bi-cubic, and generator.

As shown in Fig. 8, for all anchors, the bi-cubic method did not always result in improved results when compared with the generator method. To improve these cases, we conducted a total of nine experiments by applying three criteria to AS-RIG as shown in Table 5. Of the nine experiments, AS-RIG: SSIM

TABLE 6. Comparison of AS-RIG with SYSU-MM01 (unit: %).

Method		SYSU-MM01			
		Rank 1	Rank 10	Rank 20	mAP
Original [22]		21.29	48.52	63.34	21.25
Only bi-linear		15.36	40.43	56.87	17.29
Only bi-cubic		16.98	41.80	58.22	17.78
Only generator		18.87	40.97	59.30	19.12
AS-RIG: MSE	Threshold A	15.90	40.70	57.95	17.78
	Threshold B	**21.83**	51.48	64.69	19.70
	Threshold C	20.49	52.02	64.96	19.71
AS-RIG: SSIM	Threshold A	**21.83**	52.56	66.58	20.43
	Threshold B	21.56	**56.60**	68.46	20.83
	Threshold C	20.75	55.80	**68.73**	20.74
AS-RIG: FD	Threshold A	19.95	54.99	68.46	20.83
	Threshold B	20.49	55.80	**68.73**	**20.98**
	Threshold C	20.49	54.72	66.31	20.85

(Threshold C) showed the best performance with 57.60% rank 1, 88.96% rank 10, 94.13% rank 20, and 41.34% mAP. In addition, all nine of the experiments of AS-RIG showed better results than the non-adaptive selection methods. Based on this experimental result, it is shown that the adaptive selection method can improve the performance better than the previous method of reconstructing input data.

2) SYSU-MM01 (ABLATION STUDIES)

To verify the AS-RIG proposed in this paper, we performed the same experiment as DBPerson-Recog-DB1 using SYSU-MM01 with more variables. As in the experiment of DBPerson-Recog-DB1, we also used rank 1, rank 10, rank 20, and mAP. The experimental results for SYSU-MM01 are displayed in Table 6. As with the original in Table 5, the original [22] in Table 6 is the result of not considering the difference in resolution between the visible-light frame and thermal frame. In order to consider the actual environment, we experimented by downsizing the infrared frames of SYSU-MM01 to one eighth of the sizes. Among the non-adaptive selection methods in the first experiment, the ''only generator'' method showed the best performances of 18.87% at rank 1, 59.30% at rank 20, and 19.12% at mAP. However, the experimental result showed that the ''only bi-cubic'' method had 41.80%, which was the highest performance at rank 10.

As can be seen from these experimental results and Fig. 9, the generator method did not always have better results than the bi-cubic method for all anchors. To improve the performance of person ReID, we conducted a total of nine experiments by applying three criteria. In the experiments on DBPerson-Recog-DB1, excellent conditions existed in all evaluation categories, whereas in the case of SYSU-MM01, the conditions corresponding to the best case varied according to the experimental evaluation criteria (Rank N and mAP). In rank 1, AS-RIG: MSE (Threshold B) and AS-RIG: SSIM (Threshold A) had the highest performance at 20.83%, and

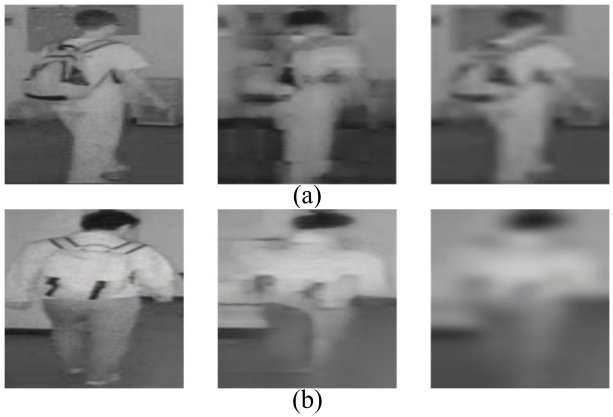


FIGURE 9. Examples of comparison bi-cubic with generator test result on SYSU-MM01: (a) The left frame is the original frame, the middle frame is recorded rank 66 (reconstructed by generator), and the right frame is recorded rank 4 (resized by bi-cubic). (b) The left frame is the original frame, the middle frame is recorded rank 6 (reconstructed by generator), and the right frame is recorded rank 141 (resized by bi-cubic).

TABLE 7. Comparisons with the state-of-the-art methods (unit: %).

Method	DBPerson-Recog-DB1		
	Rank 1	Rank 10	mAP
Original	70.93	86.39	66.04
HI-CMD [30]	66.49	84.51	52.83
AS-RIG (proposed method)	69.46	85.92	60.39
	SYSU-MM01		
Original	34.94	77.58	35.94
HI-CMD [30]	30.20	71.25	30.05
AS-RIG (proposed method)	33.29	76.98	34.33

in rank 10, AS-RIG: SSIM (Threshold B) had the best performance at 56.60%. In rank 20, AS-RIG: SSIM (Threshold C) and FD (Threshold B) showed the highest performance at 68.73%, and in mAP, AS-RIG: FD (Threshold B) showed the best performance at 20.95%.

The reason for which the method of highest performance varies for each evaluation category can be considered the characteristics of SYSU-MM01. As it is a dataset with various variables, it is possible to infer that there is an optimal criterion for each environment. However, all nine experimental results showed superiority to non-adaptive selection methods. From this, AS-RIG confirmed the potential for improving the performance better than the previous method of reconstructing input data.

3) COMPARISONS WITH STATE-OF-THE-ART METHODS

To verify the AS-RIG proposed in this paper, we performed the comparison experiment with HI-CMD [30], one of the state-of-the-art methods, using DBPerson-Recog-DB1 and SYSU-MM01.

The experimental results are displayed in Table 7. As shown in the results with original data in Tables 5 and 6, the results with original in Table 7 are those in case of using the high-resolution thermal images. HI-CMD method obtains

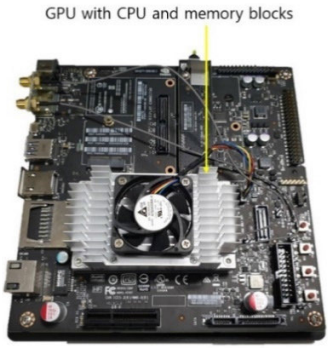


FIGURE 10. Jetson TX2 embedded system.

the high-resolution thermal image from low-resolution ones by bi-cubic method [30]. For fair comparisons, the same method of person re-identification of HI-CMD was used for all the methods. As shown in the results of Table 7, it was proved that proposed AS-RIG to obtain the high-resolution thermal image from low-resolution ones shows the higher accuracies than those by the state-of-the-art method. In addition, the accuracies of rank 1 by proposed AS-RIG are similar to those by using the original high-resolution thermal images.

4) COMPARISONS OF PROCESSING TIME

In the next experiment, the computing speed of the proposed method was compared using a desktop computer including an Intel® Core™ i7-7700K CPU @ 3.60 GHz processor (4 cores), a main memory of 32 GB, and an NVIDIA GeForce GTX 1070 (1920 CUDA cores) with a graphic processing unit (GPU) card that has a memory of 8 GB [39], and Jetson TX2 embedded system [46] as shown in Figure 10. Jetson TX2 system is equipped with NVIDIA Pascal™ GPU architecture with 256 NVIDIA CUDA cores, 8 GB 128-bit LPDDR4 memory, and dual-core NVIDIA Denver 2 64-Bit CPU. As shown in Figure 1, AS-RIG part includes the selection of the generator of CycleGAN or bi-cubic interpolation. In our experiments, it takes 9.9 ms and 22.3 ms for the generator of CycleGAN on desktop computer and Jetson TX2 embedded system, respectively. In addition, it takes 0.1 ms and 1.2 ms for the bi-cubic interpolation on desktop computer and Jetson TX2 embedded system, respectively. Because the processing time of AS-RIG in Table 8 is almost similar to the average time for the generator of CycleGAN and bi-cubic interpolation, we can estimate that the number of selection for the generator of CycleGAN is almost similar to that for the bi-cubic interpolation in our AS-RIG with all the testing images. The reason why processing time for person ReID is larger than that for AS-RIG is that the image of 224×224 pixels is used for the input to the ResNet-50 (pretrained with ImageNet database [47] and fine-tuned with our training database) [33] for person ReID.

As shown in Table 8, total processing time for one pair of input visible and thermal images are 23.8 ms and 52.7 ms on desktop computer and Jetson TX2 embedded system, respectively, which corresponds to the processing speed of 42.02 (1000/23.8) frames/sec and 18.98 (1000/52.7) frames/sec. From this results, we can confirm that our proposed method can be operated at fast speed on embedded system of limited processing power in addition to desktop computer.

TABLE 8. Comparisons of processing time of proposed method for one pair of input visible and thermal images on desktop computer and Jetson TX2 (unit: ms).

Platform	AS-RIG	ResNet-50 for person ReID	Total
Desktop computer	5.1	18.7	23.8
Jetson TX2	11.9	40.8	52.7

V. CONCLUSION

In this work, we proposed a new perspective for improving the performance of cross-modality person ReID using AS-RIG. This study was optimized for an actual environment by considering the characteristics of a thermal camera with a resolution that is lower than that of a visible-light camera. In order to improve the cross-modality person ReID using the traditional interpolation method, reconstruction by generator was applied as input data reconstruction. In addition, we compared the generator and interpolation (bi-linear and bi-cubic) methods. Based on the results of the analysis, we found that reconstructed thermal frames are not always more favorable to person ReID than reconstructed thermal frames by bi-cubic. To apply this analysis result, we proposed an adaptive selection method using MSE, SSIM, and FD. AS-RIG was evaluated using two open databases, namely DBPerson-Recog-DB1 and SYSU-MM01. Experiment results confirmed that the performance improved significantly in both databases compared to the non-adaptive selection method. There are three reasons why we did not train the system in an end-to-end manner as follows.

First, the purpose of our research is to raise an issue that the many researchers miss the difference of physical properties between visible-light and thermal cameras, and to propose a way to solve it. The difference of image resolutions between two cameras is also another challenge for person ReID research, which was not dealt with by previous works.

Second, if we train the system in an end-to-end manner for the identity information, the identity-information can highly depend on the ReID network, which are not reliable enough to train the generator of adaptive selection, and that makes the network overfitted. It is confirmed that as shown in Table 7, our method shows the higher accuracies than HI-CMD [30] which is one of the state-of-the-art methods and trains the system in an end-to-end manner.

Third, the system complexity and training time also increase in the training of end-to-end manner. Because we do not propose person ReID method itself, but mainly propose the method of adaptive selection model of generator and interpolation method. Therefore, our proposed selection model can be used for any kinds of person ReID method. That is why we do not train the system in an end-to-end manner, but separately train our selection model and ReID model.

In future work, we would perform the experiments by changing the number of testing samples, and check how much these changes affect the accuracies of person ReID. We would also have comparative experiments with the state-of-the-art methods [8], [51]–[53] as future works. In addition, we would research the method of applying proposed AS-RIG to different tasks of image super-resolution reconstruction in face or gender recognition at a distance.

REFERENCES

[1] X. Wang, "Intelligent multi-camera video surveillance: A review," *Pattern Recognit. Lett.*, vol. 34, no. 1, pp. 3–19, Jan. 2013.

[2] D. Wu, S.-J. Zheng, X.-P. Zhang, C.-A. Yuan, F. Cheng, Y. Zhao, Y.-J. Lin, Z.-Q. Zhao, Y.-L. Jiang, and D.-S. Huang, "Deep learning-based methods for person re-identification: A comprehensive review," *Neurocomputing*, vol. 337, pp. 354–371, Apr. 2019.

[3] L. Zheng, Y. Yang, and A. G. Hauptmann, "Person re-identification: Past, present and future," 2016, *arXiv:1610.02984*. [Online]. Available: http://arxiv.org/abs/1610.02984

[4] K. Wang, H. Wang, M. Liu, X. Xing, and T. Han, "Survey on person re-identification based on deep learning," *CAAI Trans. Intell. Technol.*, vol. 3, no. 4, pp. 219–227, Dec. 2018.

[5] M. O. Almasawa, L. A. Elrefaei, and K. Moria, "A survey on deep learning-based person re-identification systems," *IEEE Access*, vol. 7, pp. 175228–175247, 2019.

[6] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. H. Hoi, "Deep learning for person re-identification: A survey and outlook," 2020, *arXiv:2001.04193*. [Online]. Available: http://arxiv.org/abs/2001.04193

[7] G. Wang, T. Zhang, J. Cheng, S. Liu, Y. Yang, and Z. Hou, "RGB-infrared cross-modality person re-identification via joint pixel and feature alignment," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3623–3632.

[8] M. Ye, J. Shen, and L. Shao, "Visible-infrared person re-identification via homogeneous augmented TRI-modal learning," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 728–739, 2021.

[9] D. Nguyen, H. Hong, K. Kim, and K. Park, "Person recognition system based on a combination of body images from visible light and thermal cameras," *Sensors*, vol. 17, no. 3, p. 605, Mar. 2017.

[10] A. Wu, W.-S. Zheng, S. Gong, and J. Lai, "RGB-IR person re-identification by cross-modality similarity preservation," *Int. J. Comput. Vis.*, vol. 128, no. 6, pp. 1765–1785, Jun. 2020.

[11] T. Huang and S. Russell, "Object identification: A Bayesian context," in *Proc. 15th Int. Joint Conf. Artif. Intell.*, Nagoya, Japan, Aug. 1997, pp. 1276–1282.

[12] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2360–2367.

[13] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3908–3916.

[14] D. Chen, D. Xu, H. Li, N. Sebe, and X. Wang, "Group consistent similarity learning via deep CRF for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8649–8658.

[15] Z. Zheng, L. Zheng, and Y. Yang, "A discriminatively learned CNN embedding for person re-identification," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 14, no. 1, p. 13, Dec. 2017.

[16] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, "Person re-identification by multi-channel parts-based CNN with improved triplet loss function," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1335–1344.

[17] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," 2017, *arXiv:1703.07737*. [Online]. Available: http://arxiv.org/abs/1703.07737

[18] W. Chen, X. Chen, J. Zhang, and K. Huang, "Beyond triplet loss: A deep quadruplet network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1320–1329.

[19] A. Wu, W.-S. Zheng, and J.-H. Lai, "Robust depth-based person re-identification," *IEEE Trans. Image Process.*, vol. 26, no. 6, pp. 2588–2603, Jun. 2017.

[20] M. Ye, Z. Wang, X. Lan, and P. C. Yuen, "Visible thermal person re-identification via dual-constrained top-ranking," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 1092–1099.

[21] M. Ye, X. Lan, J. Li, and P. C. Yuen, "Hierarchical discriminative learning for visible thermal person re-identification," in *Proc. 32nd Assoc. Advancement Artif. Intell. Conf. Artif. Intell.*, New Orleans, LA, USA, Feb. 2018, pp. 7501–7508.

[22] J. K. Kang, T. M. Hoang, and K. R. Park, "Person re-identification between visible and thermal camera images based on deep residual CNN using single input," *IEEE Access*, vol. 7, pp. 57972–57984, 2019.

[23] J.-W. Lin and H. Li, "HPILN: A feature learning framework for cross-modality person re-identification," 2019, *arXiv:1906.03142*. [Online]. Available: http://arxiv.org/abs/1906.03142

[24] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 815–823.

[25] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang, "Camera style adaptation for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5157–5166.

[26] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2223–2232.

[27] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer GAN to bridge domain gap for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 79–88.

[28] Y. Ge, Z. Li, H. Zhao, G. Yin, S. Yi, X. Wang, and H. Li, "FD-GAN: Pose-guided feature distilling GAN for robust person re-identification," in *Proc. Adv. Nueral Inf. Process. Syst.*, Montreal, QC, Canada, Dec. 2018, pp. 1222–1233.

[29] Z. Zhang, S. Jiang, C. Huang, Y. Li, and R. Yi Da Xu, "RGB-IR cross-modality person ReID based on teacher-student GAN model," 2020, *arXiv:2007.07452*. [Online]. Available: http://arxiv.org/abs/2007.07452

[30] S. Choi, S. Lee, Y. Kim, T. Kim, and C. Kim, "Hi-CMD: Hierarchical cross-modality disentanglement for visible-infrared person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10257–10266.

[31] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[32] D. Salomon, *Data Compression: The Complete Reference*, 4th ed. New York, NY, USA: Springer-Verlag, 2006.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[34] *C600 Webcam Camera*. Accessed: Sep. 16, 2020. [Online]. Available: https://support.logitech.com/en_us/product/5869

[35] *Tau2 Thermal Imaging Camera*. Accessed: Sep. 16, 2020. [Online]. Available: http://www.flir.com/cores/display/?id=54717

[36] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980

[37] J. Durchi, E. Hazan, and Y. Singer, "Adaptive subgradient method for online learning and stochastic optimization," *J. Mach. Learn. Res.*, vol. 12, pp. 2121–2159, Feb. 2011.

[38] T. Tieleman and G. Hinton, "Lecture 6.5–'RMSProp: Divide the gradient by a running average of its recent magnitude," COURSERA, Neural Netw. Mach. Learn., Univ. Toronto, Toronto, ON, Canada, Tech. Rep. 4, 2012, pp. 26–31.

[39] *Geforce GTX 1070*. Accessed: Sep. 16, 2020. [Online]. Available: https://www.geforce.com/hardware/desktop-gpus/geforce-gtx-1070/specifications

[40] *mAP*. Accessed: Sep. 16, 2020. [Online]. Available: https://en.wikipedia.org/wiki/Evaluation_measures_(information_retrieval)#Mean_average_precision

[41] *AS-RIG With Algorithm*. Accessed: Sep. 16, 2020. [Online]. Available: http://dm.dgu.edu/link.html

[42] *Pytorch*. Accessed: Sep. 16, 2020. [Online]. Available: https://pytorch.org/

[43] F. Luo, H. Huang, Y. Duan, J. Liu, and Y. Liao, "Local geometric structure feature for dimensionality reduction of hyperspectral imagery," *Remote Sens*, vol. 9, no. 8, pp. 1–23, 2017.

[44] G. Shi, H. Huang, and L. Wang, "Unsupervised dimensionality reduction for hyperspectral imagery via local geometric structure feature learning," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 8, pp. 1425–1429, Aug. 2020.

[45] Z. Qin, W. He, F. Deng, M. Li, and Y. Liu, "SRPRID: Pedestrian re-identification based on super-resolution images," *IEEE Access*, vol. 7, pp. 152891–152899, 2019.
[46] *Jetson TX2 System*. Accessed: Sep. 16, 2020. [Online]. Available: https://www.nvidia.com/en-us/autonomous-machines/embedded-systems-dev-kits-modules/
[47] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2012, pp. 1097–1105.
[48] X. Li, W.-S. Zheng, X. Wang, T. Xiang, and S. Gong, "Multi-scale learning for low-resolution person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3765–3773.
[49] Z. Wang, R. Hu, J. Jiang, J. Jiang, C. Liang, and J. Wang, "Scale-adaptive low-resolution person re-identification via learning a discriminating surface," in *Proc. Int. Joint Conf. Artif. Intell.*, Jul. 2016, pp. 2669–2675.
[50] X.-Y. Jing, X. Zhu, F. Wu, R. Hu, X. You, Y. Wang, H. Feng, and J.-Y. Yang, "Super-resolution person re-identification with semi-coupled low-rank discriminant dictionary learning," *IEEE Trans. Image Process.*, vol. 26, no. 3, pp. 1363–1378, Mar. 2017.
[51] M. Ye, J. Shen, D. J. Crandall, L. Shao, and J. Luo, "Dynamic dual-attentive aggregation learning for visible-infrared person re-identification," in *Proc. Eur. Conf. Comput. Vis.*, Aug. 2020, pp. 229–247.
[52] M. Ye, X. Lan, Q. Leng, and J. Shen, "Cross-modality person re-identification via modality-aware collaborative ensemble learning," *IEEE Trans. Image Process.*, vol. 29, pp. 9387–9399, Jun. 2020.
[53] M. Ye, X. Lan, Z. Wang, and P. C. Yuen, "Bi-directional center-constrained top-ranking for visible thermal person re-identification," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 407–419, Jun. 2020.
[54] M. Ye, J. Shen, X. Zhang, P. C. Yuen, and S.-F. Chang, "Augmentation invariant and instance spreading feature for softmax embedding," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Aug. 3, 2020, doi: 10.1109/TPAMI.2020.3013379.

JIN KYU KANG received the B.S. degree in electronics and electrical engineering from Dongguk University, Seoul, South Korea, in 2016, where he is currently pursuing the combined M.S. and Ph.D. degree in electronics and electrical engineering. His research interests include biometrics and deep learning. He implemented the overall system and wrote the draft of the original article.

MIN BEOM LEE received the B.S. degree in information and telecommunication engineering from Dongyang Mirae University, Seoul, South Korea, in 2016. He is currently pursuing the combined M.S. and Ph.D. degree in electronics and electrical engineering from Dongguk University. His research interests include biometrics and deep learning. He assisted with performing the experiments and analyzing the results.

HYO SIK YOON received the B.S. degree in electronics engineering from Kangwon National University, Chuncheon, South Korea, in 2015. He is currently pursuing the combined M.S. and Ph.D. degree in electronics and electrical engineering with Dongguk University. His research interests include pattern recognition and deep learning. He assisted with performing the experiments and analyzing the results.

KANG RYOUNG PARK (Member, IEEE) received the B.S. and M.S. degrees in electronic engineering and the Ph.D. degree in electrical and computer engineering from Yonsei University, Seoul, South Korea, in 1994, 1996, and 2000, respectively. He has been a Professor with the Division of Electronics and Electrical Engineering with Dongguk University, since March 2013. His research interests include image processing and deep learning. He supervised this research and assisted with the revision of the draft of the original article.