ThermalGAN: Multimodal Color-to-Thermal Image Translation for Person Re-identification in Multispectral Dataset

Vladimir V. Kniaz[1,2(✉)], Vladimir A. Knyaz[1,2], Jiří Hladůvka[3],
Walter G. Kropatsch[3], and Vladimir Mizginov[1]

[1] State Research Institute of Aviation Systems (GosNIIAS), Moscow, Russia
{vl.kniaz,knyaz,vl.mizginov}@gosniias.ru
[2] Moscow Institute of Physics and Technology (MIPT), Dolgoprudny, Russia
[3] PRIP, Institute of Visual Computing and Human-Centered Technology,
Vienna, Austria
{jiri,krw}@prip.tuwien.ac.at

Abstract. We propose a ThermalGAN framework for cross-modality color-thermal person re-identification (ReID). We use a stack of generative adversarial networks (GAN) to translate a single color probe image to a multimodal thermal probe set. We use thermal histograms and feature descriptors as a thermal signature. We collected a large-scale multispectral ThermalWorld dataset for extensive training of our GAN model. In total the dataset includes 20216 color-thermal image pairs, 516 person ID, and ground truth pixel-level object annotations. We made the dataset freely available (http://www.zefirus.org/ThermalGAN/). We evaluate our framework on the ThermalWorld dataset to show that it delivers robust matching that competes and surpasses the state-of-the-art in cross-modality color-thermal ReID.

Keywords: Person re-identification · Conditional GAN
Thermal images

1 Introduction

Person re-identification (ReID) is of primary importance for tasks such as video surveillance and photo-tagging. It has been the focus of intense research in recent years. While modern methods provide excellent results during in a well-lit environment, their performance is not robust without a suitable light source.

Infrared cameras perceive thermal emission that is invariant to the lighting conditions. However, challenges of cross-modality color-infrared matching

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-11024-6_46) contains supplementary material, which is available to authorized users.

L. Leal-Taixé and S. Roth (Eds.): ECCV 2018 Workshops, LNCS 11134, pp. 606–624, 2019.
https://doi.org/10.1007/978-3-030-11024-6_46

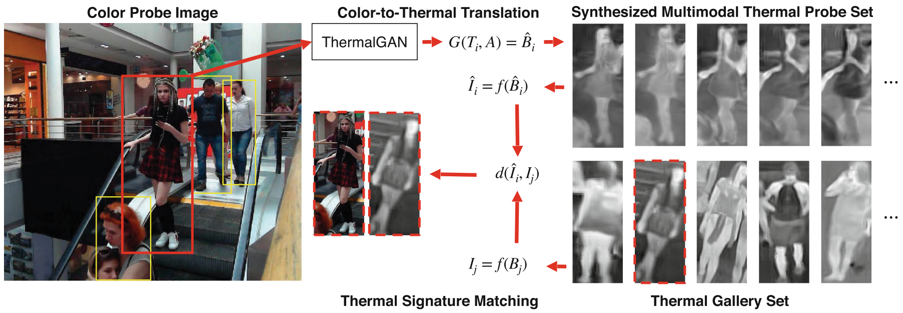
Color Probe Image
Color-to-Thermal Translation
Synthesized Multimodal Thermal Probe Set
ThermalGAN
$G(T_i, A) = \hat{B}_i$
$\hat{I}_i = f(\hat{B}_i)$
...
$d(\hat{I}_i, I_j)$
...
$I_j = f(B_j)$
Thermal Signature Matching
Thermal Gallery Set

Fig. 1. Overview of color-thermal ReID using our `ThermalGAN` framework. We transform a single color probe image A to multimodal thermal probe set $\{B_1, \ldots, B_i\}$. We use thermal signatures I to perform matching with real thermal gallery set. (Color figure online)

reduce benefits of night mode infrared cameras. Recently cross-modality color-to-thermal matching received a lot of scholar attention [35,37,51,52]. Multiple datasets with infrared images [33,35,37,47] were developed for cross-modality infrared-to-color person ReID. Thermal cameras operate in longwave infrared (LWIR, 8–14 μ) and provide real temperatures of a person body which are more stable to viewpoint changes than near-infrared images [45,54,58].

This paper is focused on the development of a `ThermalGAN` framework for color-thermal cross-modality person ReID. We use assumptions of Bhuiyan [4] and Zhu [62] as a starting point for our research to develop a color-to-thermal transfer framework for cross-modality person ReID. We perform matching using calibrated thermal images to benefit from the stability of surface temperatures to changes in light intensity and viewpoint. Matching is performed in two steps. Firstly, we model a person appearance in a thermal image conditioned by a color image. We generate a multimodal thermal probe set from a single color probe image using a generative adversarial network (GAN). Secondly, we perform matching in thermal images using the synthesized thermal probe set and a real thermal gallery set (Fig. 1).

We collected a new ThermalWorld dataset to train our GAN framework and to test the ReID performance. The dataset contains two parts: ReID and Visual Objects in Context (VOC). The ReID split includes 15118 pairs of color and thermal images and 516 person ID. The VOC part is designed for training color-to-thermal translation using a GAN framework. It consists of 5098 pairs of color and thermal images and pixel-level annotations for ten classes: person, car, truck, van, bus, building, cat, dog, tram, boat.

We perform an evaluation of baselines and our framework on the ThermalWorld dataset. The results of the evaluation are encouraging and show that our `ThermalGAN` framework achieves and surpasses the state-of-the-art in the cross-modality color-thermal ReID. The new `ThermalGAN` framework will be able to perform matching of color probe image with thermal gallery set in video surveillance applications.

Section 2 presents the structure of the developed ThermalWorld dataset. In Sect. 3 we describe the `ThermalGAN` framework and thermal signature-based matching. In Sect. 4 we give an evaluation of ReID baselines and the `ThermalGAN` framework on the ThermalWorld dataset.

1.1 Contributions

We present three main contributions: (1) the `ThermalGAN` framework for color-to-thermal image translation and ReID, (2) a large-scale multispectral ThermalWorld dataset with two splits: ReID with 15118 color-thermal image pairs and 516 person ID, and VOC with 5098 pairs color-thermal image pairs with ground truth pixel-level object annotations of ten object classes, (3) an evaluation of the modern cross-modality ReID methods on ThermalWorld ReID dataset.

2 Related Work

2.1 Person Re-identification

Person re-identification has been intensively studied by computer vision society recently [3,4,9,12,40,47]. While new methods improve the matching performance steadily, video surveillance applications still pose challenges for ReID systems. Recent methods regarding person ReID can be divided into three kinds of approaches [4]: direct methods, metric learning methods and transform learning methods.

In [4] an overview of modern ReID methods was performed, and a new transform learning-based method was proposed. The method models an appearance of a person in a new camera using cumulative weight brightness transfer function (CWBTF). The method leverages a robust segmentation technique to segment the human image into meaningful parts. Matching of features extracted only from the body area provides an increased ReID performance. The method also exploits multiple pedestrian detections to improve the matching rate.

While the method provides the state-of-the-art performance on color images, night-time application requires additional modalities to perform robust matching in low-light conditions. Cross-modality color-infrared matching is gaining increasing attention. Multiple multispectral datasets were collected in recent years [33,35,37,47,51]. SYSU-MM01 dataset [47] includes unpaired color and near-infrared images. RegDB dataset [51] presents color and infrared images for evaluation of cross-modality ReID methods. Evaluation of modern methods on these datasets has shown that color-infrared matching is challenging. Nevertheless, it provides an increase in ReID robustness during the night-time.

Thermal camera has received a lot of scholar attention in the field of video surveillance [8,53]. While thermal cameras provide a significant boost in pedestrian detection [42,50] and ReID with paired color and thermal images [33], cross-modality person ReID is challenging [33–37] due to severe changes in a person appearance in color and thermal images.

Recently, generative adversarial networks (GAN) [13] have demonstrated encouraging results in arbitrary image-to-image translation applications. We hypothesize that color-to-thermal image translation using a dedicated GAN framework can increase color-thermal ReID performance.

2.2 Color-to-Thermal Translation

Transformation of the spectral range of an image has been intensively studied in such fields as transfer learning [39,46] domain adaptation [23–25,32] and cross-domain recognition [1,17,19,21,47,49,54,55]. In [30] a deep convolutional neural network (CNN) was proposed for translation of a near-infrared image to a color image. The approach was similar to image colorization [15,56] and style transfer [27,44] problems that were actively studied in recent years. The proposed architecture succeeded in a translation of near-infrared images to color images. Transformation of LWIR images is more challenging due to the low correlation between the red channel of a color image and a thermal image.

2.3 Generative Adversarial Networks

GANs increased the quality of image-to-image translation significantly [20,54,58] using an antagonistic game approach [13]. Isola et al. [20] proposed a `pix2pix` GAN framework for arbitrary image transformations using geometrically aligned image pairs sampled from source and target domains. The framework succeeded in performing arbitrary image-to-image translations such as season change and object transfiguration. Zhang et al. [54,58] trained the `pix2pix` network to transform a thermal image of a human face to the color image. The translation improved the quality of a face recognition performance in a cross-modality thermal to visible range setting. While human face has a relatively stable temperature, color-thermal image translation for the whole human body with an arbitrary background is more ambiguous and conditioned by the sequence of events that have occurred with a person.

We hypothesize that multimodal image generation methods can model multiple possible thermal outputs for a single color probe image. Such modeling can improve the ReID performance. Zhu et al. proposed a `BicycleGAN` framework [63] for modeling a distribution of possible outputs in a conditional generative modeling setting. To resolve the ambiguity of the mapping Zhu et al. used a randomly sampled low-dimension latent vector. The latent vector is produced by an encoder network from the generated image and compared to the original latent vector to provide an additional consistency loss. We propose a conditional color-to-thermal translation framework for modeling of a set of possible person appearances in a thermal image conditioned by a single color image.

3 ThermalWorld Dataset

We collected a new ThermalWorld dataset to train and evaluate our cross-modality ReID framework. The dataset was collected with multiple FLIR ONE

PRO cameras and divided into two splits: ReID and Visual Objects in Context (VOC). The ReID split includes 15118 aligned color and thermal image pairs of 516 IDs. The VOC split was created for effective color-to-thermal translation GAN training. It contains 5098 color and thermal image pairs and pixel-level annotations of ten object classes: person, car, truck, van, bus, building, cat, dog, tram, boat.

Initially, we have tried to train a color-to-thermal translation GAN model using only the ReID split. However, the trained network has poor generalization ability due to a limited number of object classes and backgrounds. This stimulated us to collect a large-scale dataset with aligned pairs of color and thermal images. The rest of this section presents a brief dataset overview. For more details on the dataset, please refer to the supplementary material.

3.1 ThermalWorld ReID Split

The ReID split includes pairs of color and thermal images captured by sixteen FLIR ONE PRO cameras. Sample images from the dataset are presented in Fig. 2. All cameras were located in a shopping mall area. Cameras #2, 9, 13 are located in underground passages with low-light conditions. Cameras #1, 3, 7, 8, 10, 12, 14 are located at the entrances and present both day-time and night-time images. Cameras #15,16 are placed in the garden. The rest of the cameras are located inside the mall.

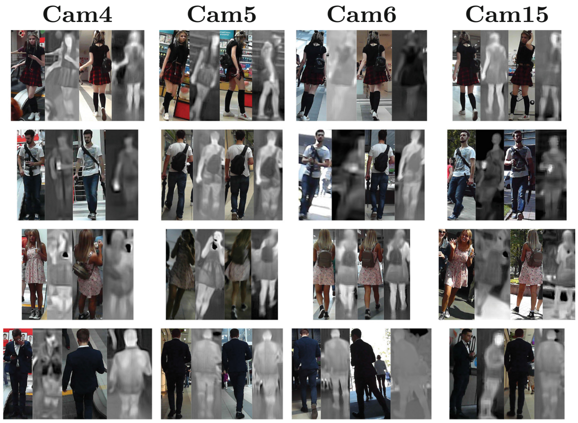

Fig. 2. Examples of person images from ThermalWorld ReID dataset.

We have developed a dedicated application for a smartphone to record sequences of thermal images using FLIR ONE PRO. Comparison to previous ReID datasets is presented in Table 1.

Table 1. Comparison to previous ReID datasets. #/# represents the number of color/infrared images or cameras.

Dataset	#ID	#images	#cam	Color	NIR	Thermal
iLDS [61]	119	476	2	✓	×	×
CAVIAR [6]	72	610	2	✓	×	×
PRID2011 [18]	200	971	2	✓	×	×
VIPER [14]	632	1264	2	✓	×	×
CUHK01 [28]	972	1942	2	✓	×	×
CUHK03 [29]	1467	13164	6	✓	×	×
SYSU [16]	502	24448	2	✓	×	×
Market [60]	1501	32668	6	✓	×	×
MARS [59]	1261	1191003	6	✓	×	×
RegDB [37]	412	4120/4120	1/1	✓	×	✓
SYSU-MM01 [47]	491	287628/15792	4/2	✓	✓	×
ThermalWorld	516	15818/15818	16/16	✓	×	✓

3.2 ThermalWorld VOC Split

The VOC split of the dataset was collected using two FLIR ONE PRO cameras. We use insights of developers of previous multispectral datasets [2,8,11,19,43,55] and provide new object classes with pixel-level annotations. The images were collected in different cities (Paris, Strasbourg, Riva del Garda, Venice) during all seasons and in different weather conditions (sunny, rain, snow). Captured object temperatures range from −20 °C to +40 °C.

We hypothesized that training a GAN to predict the relative temperature contrasts of an object (e.g., clothes/skin) instead of its absolute temperature can improve the translation quality. We were inspired by the previous work on the explicit encoding of multiple modes in the output [63], and we assumed that the thermal segmentation that provides average temperatures of the emitting objects in the scene could resolve the ambiguity of the generated thermal images. Examples from ThermalWorld VOC dataset are presented in Fig. 3.

We manually annotated the dataset, to automatically extract an object's temperature from the thermal images. Comparison to previous multispectral datasets and examples of all classes are presented in the supplementary material.

4 Method

Color-to-thermal image translation is challenging because there are multiple possible thermal outputs for a single color input. For example, a person in a cold autumn day and a hot summer afternoon may have a similar appearance in the visible range, but the skin temperature will be different.

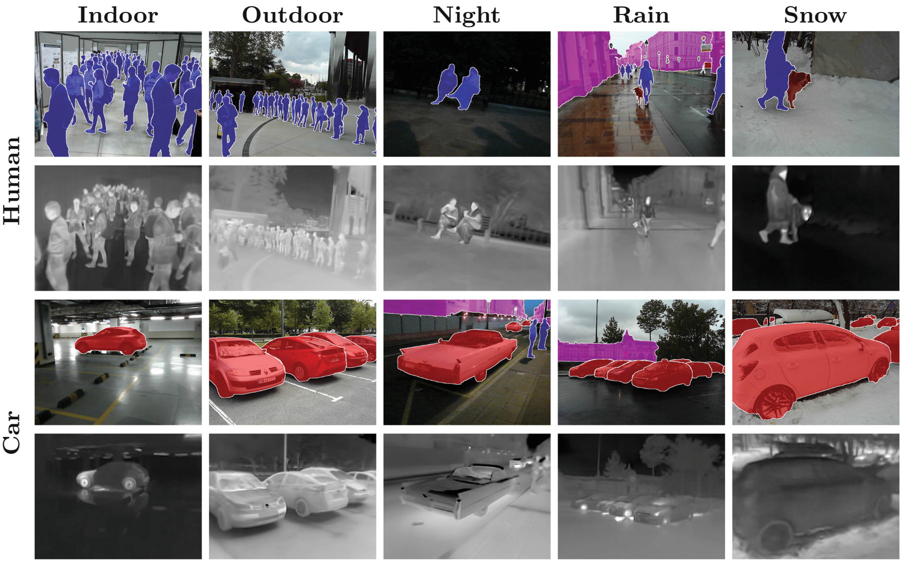

Fig. 3. Examples of annotated images in ThermalWorld VOC dataset.

We have experimented with multiple state-of-the-art GAN frameworks [5, 20, 26, 63] for multi-modal image translation on the color-to-thermal task. We have found that GANs can predict object temperature with accuracy of approximately 5 °C.

However, thermal images must have accuracy of 1 °C to make local body temperature contrasts (e.g., skin/cloth) distinguishable. To improve the translation quality we developed two-step approach inspired by [48]. We have observed that relative thermal contrasts (e.g., eyes/brow) are nearly invariant to changes in the average body temperature due to different weather conditions.

We hypothesize that a prediction of relative thermal contrasts can be conditioned by an input color image and average temperatures for each object. Thus, we predict an absolute object temperature in two steps (Fig. 4). Firstly, we predict average object temperatures from an input color image. We term the resulting image as a "thermal segmentation" image $\hat{S}$.

Secondly, we predict the relative local temperature contrasts $\hat{R}$, conditioned by a color image A and a thermal segmentation $\hat{S}$. The sum of a thermal segmentation and temperature contrasts provides the thermal image: $\hat{B} = \hat{S} + \hat{R}$.

The sequential thermal image synthesis has two advantages. Firstly, the problem remains multimodal only in the first step (generation of thermal segmentation). Secondly, the quality of thermal contrasts prediction is increased due to lower standard deviation and reduced range of temperatures.

To address the multimodality in color-to-thermal translation, we use a modified `BicyleGAN` framework [63] to synthesize multiple color segmentation images for a single color image. Instead of a random noise sample, we use a temperature vector T_i, which contains the desired background and object temperatures.

Kniaz2019

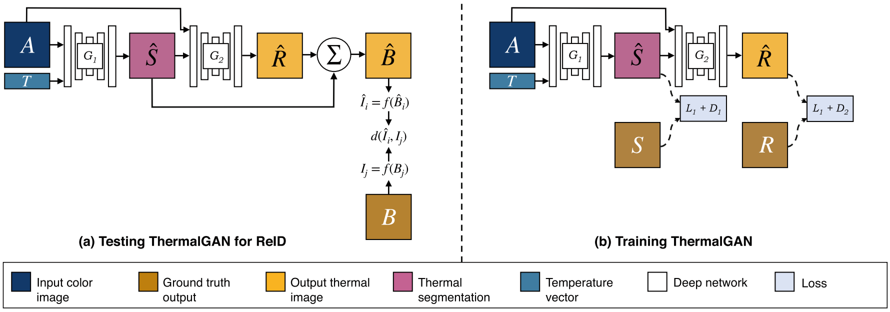

Fig. 4. Overview of our `ThermalGAN` framework.

The rest of this section presents a brief introduction to conditional adversarial networks, details on the developed `ThermalGAN` framework and features used for thermal signature matching.

4.1 Conditional Adversarial Networks

Generative adversarial networks produce an image $\hat{B}$ for a given random noise vector z, $G : z \rightarrow \hat{B}$ [13,20]. Conditional GAN receives extra bits of information A in addition to the vector z, $G : \{A, z\} \rightarrow \hat{B}$. Usually, A is an image that is transformed by the generator network G. The discriminator network D is trained to distinguish "real" images from target domain B from the "fakes" $\hat{B}$ produced by the generator. Both networks are trained simultaneously. Discriminator provides the adversarial loss that enforces the generator to produce "fakes" $\hat{B}$ that cannot be distinguished from "real" images B.

We train two GAN models. The first generator $G_1 : \{T_i, A\} \rightarrow \hat{S}_i$ synthesizes multiple thermal segmentation images $\hat{S}_i \in \mathbb{R}^{W \times H}$ conditioned by a temperature vector T_i and a color image $A \in \mathbb{R}^{W \times H \times 3}$. The second generator $G_2 : \{\hat{S}_i, A\} \rightarrow \hat{R}_i$ synthesizes thermal contrasts $\hat{R}_i \in \mathbb{R}^{W \times H}$ conditioned by a thermal segmentation $\hat{S}_i$ and the input color image A. We can produce multiple realistic thermal outputs for a single color image by changing only the temperature vector T_i.

4.2 Thermal Segmentation Generator

We use the modified `BicycleGAN` framework for thermal segmentation generator G_1. Our contribution to the original U-Net generator [41] is an addition of one convolutional layer and one deconvolutional layer to increase the output resolution. We use average background temperatures instead of the random noise sample to be able to control the appearance of the thermal segmentation. The loss function for the generator G_1 is given by [63]:

$$\begin{aligned} G_1^*(G_1, D_1) = \arg \min_{G_1} \max_{D_1} \mathcal{L}_{GAN}^{VAE}(G_1, D_1, E) + \lambda \mathcal{L}_1(G_1, E) \\ + \mathcal{L}_{GAN}(G_1, D_1) + \lambda_{\text{thermal}} \mathcal{L}_1^{\text{thermal}}(G_1, E) + \lambda_{KL} \mathcal{L}_{KL}(E), \quad (1) \end{aligned}$$

where $\mathcal{L}_{GAN}^{VAE}$ – is Variational Autoencoder-based loss [63] that stimulates the output to be multimodal, $\mathcal{L}_1$ – is an L^1 loss, $\mathcal{L}_{GAN}$ – loss provided by the discriminator D_1, $\mathcal{L}_1^{\text{thermal}}$ – is a loss of the latent temperature domain, $\mathcal{L}_{KL}$ – Kullback–Leibler-divergence loss, E – encoder network, λ – weight parameters. We train both generators independently.

4.3 Relative Thermal Contrast Generator

We hypothesize that the distribution of relative thermal contrasts conditioned by a thermal segmentation and a color image is unimodal (compare images B and R for various background temperatures in Fig. 5). Hence, we use a unimodal `pix2pix` framework [20] as a starting point for our relative contrast generator G_2. Our contribution to the original framework is two-fold. Firstly, we made the same modifications to the generator G_2 as for the generator G_1. Secondly, we use four channel input tensor. First three channels are RGB channels of an image A, the fourth channel is thermal segmentation $\hat{S}_i$ produced by generator G_1. We sum the outputs from the generators to obtain an absolute temperature image.

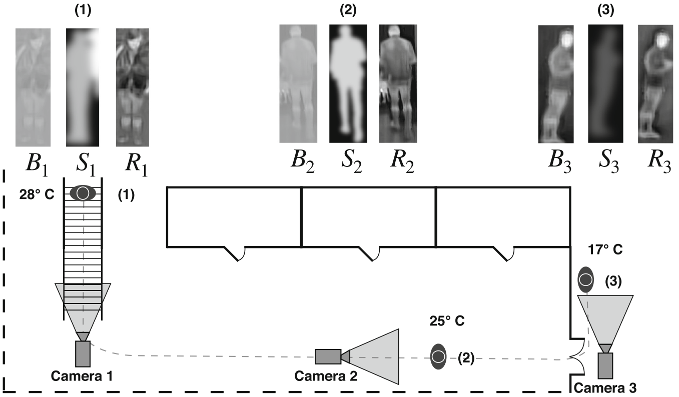

Fig. 5. Comparison of relative contrast R and absolute temperature B images for various camera locations. The relative contrast image R is equal to the difference of an absolute temperature B and a thermal segmentation S. Note that the person appearance is invariant to background temperature in relative contrast images.

4.4 Thermal Signature Matching

We use an approach similar to [4] to extract discriminative features from thermal images. The original feature signature is extracted in three steps [4]: (1) a person appearance is transferred from camera C_k to camera C_l, (2) the body region is separated from the background using stel component analysis (SCA) [22], (3) feature signature is extracted using color histograms [9] and maximally stable color regions (MSCR) [10].

However, thermal images contain only a single channel. We modify color features for absolute temperature domain. We use the monochrome ancestor of MSCR – maximally stable extremal regions (MSER) [31]. We use temperature histograms instead of color histograms. The resulting matching method includes four steps. Firstly, we transform the person appearance from a single color probe image A to multiple thermal images $\hat{B}_i$ using the `ThermalGAN` framework. Various images model possible variations of temperature from camera to camera. Unlike the original approach [4], we do not train the method to transfer person a appearance from camera to camera. Secondly, we extract body regions using SCA from real thermal images B_j in the gallery set and synthesized images $\hat{B}_i$. After that, we extract thermal signatures I from body regions

$$I = f(B) = \Big[H_t(B), f_{\text{MSER}}(B)\Big], \tag{2}$$

where H_t is a histogram of body temperatures, f_{MSER} is MSER blobs for an image B.

Finally, we compute a distance between two signatures using Bhattacharyya distance for temperature histograms and MSER distance [6,31].

$$\begin{aligned} d(\hat{I}_i, I_j) = \beta_H \cdot d_H(H_t(\hat{B}_i), H_t(B_j)) \\ + (1 - \beta_H) \cdot d_{\text{MSER}}(f_{\text{MSER}}(\hat{B}_i), f_{\text{MSER}}(B_j)), \end{aligned} \tag{3}$$

where d_H is a Bhattacharyya distance, d_{MSER} is a MSER distance [31], and β_H is a calibration weight parameter.

5 Evaluation

5.1 Network Training

The `ThermalGAN` framework was trained on the VOC split of the ThermalWorld dataset using the PyTorch library [38]. The VOC split includes indoor and outdoor scenes to avoid domain shift. The training was performed using the NVIDIA 1080 Ti GPU and took 76 h for G_1, D_1 and 68 h for G_2, D_2. For network optimization, we use minibatch SGD with an Adam solver. We set learning rate to 0.0002 with momentum parameters $\beta_1 = 0.5$, $\beta_2 = 0.999$ similar to [20].

5.2 Color-to-Thermal Translation

Qualitative Comparison. For a qualitative comparison of the ThermalGAN model on the color-to-thermal translation, we generate multiple thermal images from the independent ReID split of ThermalWorld dataset. Our goal is to keep the resulting images both realistic and diverse in terms of person relative thermal contrasts. We compare our framework with five baselines: pix2pix+noise [20], cLR-GAN [5,63], cVAE-GAN [26,63], cVAE-GAN++ [26,63], BicycleGAN [63]. All baselines were trained to convert color image to grayscale image representing perceptual thermal contrasts (8-bit, grayscale). Our ThermalGAN framework was trained to produce thermal images in degree Celsius. For comparison, they were converted to perceptual thermal intensities. Results of multimodal thermal image generation are presented in Fig. 6.

The results of pix2pix+noise are unrealistic and do not provide a changes of thermal contrast. cLR-GAN and cVAE-GAN produce a slight variation of thermal contrasts but do not translate meaningful features for ReID. cVAE-GAN++ and BicycleGAN produce a diverse output, which fails to model thermal features present in real images. Our ThermalGAN framework combines the power of BicycleGAN method with two-step sequential translation to produce the output that is both realistic and diverse.

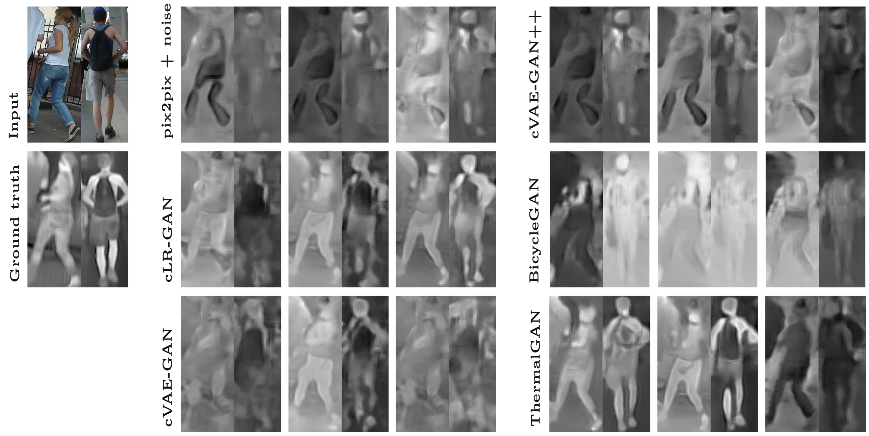

Fig. 6. Qualitative method comparison. We compare performance of various multimodal image translation frameworks on ThermalWorld ReID dataset. For each model, we present three random output. The output of ThermalGAN framework is realistic, diverse, and shows the small temperatures contrasts that are important for robust ReID. Please note that only ThermalGAN framework produces output as calibrated temperatures that can be used for thermal signature matching.

Quantitate Evaluation. We use the generated images to perform a quantitative analysis of our ThermalGAN framework and the baselines. We measure two characteristics: diversity and perceptual realism. To measure multimodal reconstruction diversity, we use the averaged Learned Perceptual Image Patch Similarity (LPIPS [57]) distance as proposed in [57,63]. For each baseline and our method, we calculate the average distance between 1600 pairs of random output thermal images, conditioned by 100 input color images. We measure perceptual realism of the synthesized thermal images using the human experts utilizing an approach similar to [56]. Real and synthesized thermal images are presented to human experts in a random order for one second. Each expert must indicate if the image is real or not. We perform the test on Amazon Mechanical Turk (AMT). We summarize the results of the quantitative evaluation in Fig. 7 and Table 2. Results of `cLR-GAN`, `BicycleGAN` and our `ThermalGAN` framework were most realistic. Our `ThermalGAN` model outperforms baselines in terms of both diversity and perceptual realism.

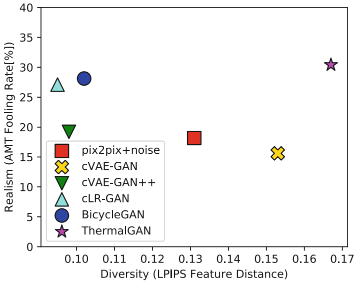

Fig. 7. Realism vs diversity for synthesized thermal images.

Table 2. Comparison with state-of-the-art multimodal image-to-image translation methods.

	Realism	Diversity
Method	AMT fooling rate [%]	LPIPS distance
Random real images	50.0%	
`pix2pix+noise` [20]	18.17	0.131
`cVAE-GAN` [26,63]	15.61	0.153
`cVAE-GAN++` [26,63]	19.21	0.098
`cLR-GAN` [5,63]	27.10	0.095
`BicycleGAN` [63]	28.12	0.102
`ThermalGAN`	30.41	0.167

5.3 ReID Evaluation Protocol

We use 516 ID from the ReID split for testing the ReID performance. Please note, that we use independent VOC split for training color-to-thermal translation. We use cumulative matching characteristic (CMC) curves and normalized area-under-curve (nAUC) as a performance measure. The CMC curve presents a recognition performance versus re-identification ranking score. nAUC is an integral score of a ReID performance of a given method. To keep our evaluation protocol consistent with related work [4], we use 5 pedestrians in the validation set. We also keep independent the gallery set and the probe set according to the common practice.

We use images from color cameras for a probe set and images from thermal cameras for a gallery set. We exclude images from cameras #2, 9, 13 from the probe set, because they do not provide meaningful data in the visible range. We use a single color image in the single-shot setting. `ThermalGAN` ReID framework uses this single color image to generate 16 various thermal images. Baseline methods use the single input color image according to the common practice.

For the multi-shot setting, we use ten color images for the probe set. Therefore ThermalGAN framework generates 16 thermal images for each color probe image and generates 160 thermal images for the probe set.

5.4 Cross-Modality ReID Baselines

We compare our framework with six baseline models including hand-crafted features HOG [7] and modern deep-learning based cross-modality matching methods: One Stream Network (OSN) [47], Two Stream Network (TSN) [47], Deep Zero-Padding (DZP) [47], Two-stream CNN network (TONE_1) [52], and Modified two-stream CNN network (TONE_2) [51].

5.5 Comparison and Analysis

We show results of a comparison of our framework and baselines on ThermalWorld ReID datasets in Table 3 for a single-shot setting and in Table 4 for the multi-shot setting. Results are given in terms of top-ranked matching rate and nAUC. We present the results in terms of CMC curves in Fig. 8.

Table 3. Experiments on ThermalWorld ReID dataset in single-shot setting.

Methods	ThermalWorld ReID single-shot					
	r = 1	r = 5	r = 10	r = 15	r = 20	nAUC
TONE_2 [51]	15.10	29.26	38.95	42.40	44.48	37.98
TONE_1 [52]	8.87	13.71	21.27	27.48	31.86	23.64
HOG [7]	14.29	23.56	33.45	40.21	43.92	34.86
TSN [47]	3.59	5.13	8.85	13.97	18.56	12.25
OSN [47]	13.29	23.11	33.05	40.06	42.76	34.27
DZP [47]	15.37	22.53	30.81	36.80	39.99	32.28
ThermalGAN	**19.48**	**33.76**	**42.69**	**46.29**	**48.19**	**41.84**

We make the following observations from the single-shot evaluation. Firstly, the two-stream network [47] performs the worst among other baselines. We assume that the reason is that fine-tuning of the network from near-infrared data to thermal range is not sufficient for effective matching. Secondly, hand-crafted HOG [7] descriptor provided discriminative features that present both in color and thermal images and can compete with some of modern methods. Finally, our ThermalGAN ReID framework succeeds in the realistic translation of meaningful features from color to thermal images and provides discriminative features for effective color-to-thermal matching.

Results in the multi-shot setting are encouraging and prove that multiple person detection improves the matching rate with a cross-modality setup. We

Table 4. Experiments on ThermalWorld ReID dataset in multi-shot setting.

Methods	ThermalWorld ReID multi-shot					
	r = 1	r = 5	r = 10	r = 15	r = 20	nAUC
TONE_2 [51]	20.11	38.19	51.62	56.73	59.38	50.30
TONE_1 [52]	11.10	17.79	24.18	31.58	36.66	27.46
HOG [7]	16.08	27.10	40.10	48.64	51.41	40.82
TSN [47]	8.71	14.97	21.10	26.30	29.87	23.21
OSN [47]	15.36	25.17	39.14	47.65	50.04	39.85
DZP [47]	14.62	24.14	33.09	39.57	44.08	34.78
ThermalGAN	**22.59**	**48.24**	**59.40**	**62.85**	**66.12**	**57.35**

conclude the following observations from the results presented in Table 4 and Fig. 8. Firstly, the performance of deep-learning-based baselines is improved in average in 5%. Secondly, multi-shot setting improves rank-5 and rank-10 recognition rates. Finally, our ThermalGAN method benefits from the multi-shot setting and can be used effectively with multiple person images provided by robust pedestrian detectors for thermal images [50].

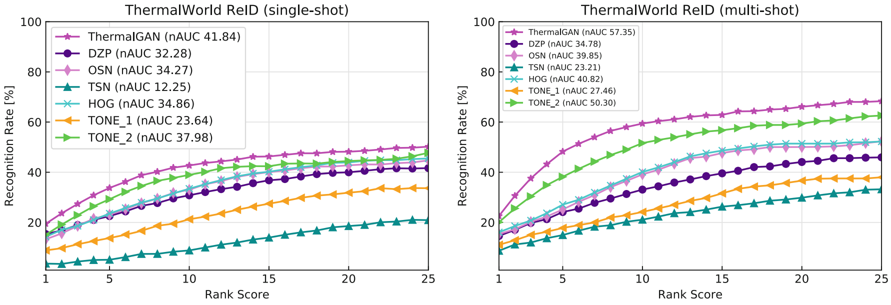

Fig. 8. CMC plot and nAUC for evaluation of baselines and ThermalGAN method in single-shot setting (left) and multi-shot setting (right).

6 Conclusion

We showed that conditional generative adversarial networks are effective for cross-modality prediction of a person appearance in thermal image conditioned by a probe color image. Furthermore, discriminative features can be extracted from real and synthesized thermal images for effective matching of thermal signatures. Our main observation is that thermal cameras coupled with a GAN ReID framework can significantly improve the ReID performance in low-light conditions.

We developed a ThermalGAN framework for cross-modality person ReID in the visible range and LWIR images. We have collected a large-scale multispectral ThermalWorld dataset to train our framework and compare it to baselines. We made the dataset publicly available. Evaluation of modern cross-modality ReID methods and our framework proved that our ThermalGAN method achieves the state-of-the-art and outperforms it in the cross-modality color-thermal ReID.

Acknowledgements. The reported study was funded by the Russian Science Foundation (RSF) according to the research project N° 16-11-00082.

References

1. Berg, A., Ahlberg, J., Felsberg, M.: A thermal infrared dataset for evaluation of short-term tracking methods. In: Swedish Symposium on Image Analysis (2015)
2. Berg, A., Ahlberg, J., Felsberg, M.: A thermal object tracking benchmark. In: 2015 12th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 1–6 (2015). http://ieeexplore.ieee.org/document/7301772/
3. Bhuiyan, A., Perina, A., Murino, V.: Person re-identification by discriminatively selecting parts and features. In: Agapito, L., Bronstein, M.M., Rother, C. (eds.) ECCV 2014. LNCS, vol. 8927, pp. 147–161. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-16199-0_11
4. Bhuiyan, A., Perina, A., Murino, V.: Exploiting multiple detections for person re-identification. J. Imaging **4**(2), 28 (2018)
5. Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., Abbeel, P.: Infogan: interpretable representation learning by information maximizing generative adversarial nets. In: Advances in Neural Information Processing Systems, pp. 2172–2180 (2016)
6. Cheng, D.S., Cristani, M., Stoppa, M., Bazzani, L., Murino, V.: Custom pictorial structures for re-identification. In: Proceedings of the British Machine Vision Conference, BMVC 2011. Universita degli Studi di Verona, Verona, Italy, January 2011
7. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005, vol. 1, pp. 886–893. IEEE (2005)
8. Davis, J.W., Keck, M.A.: A two-stage template approach to person detection in thermal imagery. In: Seventh IEEE Workshops on Application of Computer Vision, WACV/MOTIONS 2005, vol. 1, pp. 364–369. IEEE (2005)
9. Farenzena, M., Bazzani, L., Perina, A., Murino, V., Cristani, M.: Person re-identification by symmetry-driven accumulation of local features. In: 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2360–2367. IEEE, March 2010
10. Forssén, P.E.: Maximally stable colour regions for recognition and matching. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2007, pp. 1–8. IEEE (2007)
11. Généreux, F., et al.: On the figure of merit of uncooled bolometers fabricated at INO. In: Infrared Technology and Applications XLII, vol. 9819, p. 98191U. International Society for Optics and Photonics (2016)

12. Gong, S., Cristani, M., Yan, S.: Person Re-Identification (Advances in Computer Vision and Pattern Recognition). Springer, London (2014). https://doi.org/10.1007/978-1-4471-6296-4
13. Goodfellow, I., et al.: Generative adversarial nets. In: Advances in Neural Information Processing Systems, pp. 2672–2680 (2014)
14. Gray, D., Brennan, S., Tao, H.: Evaluating appearance models for recognition, reacquisition, and tracking. In: IEEE International Workshop on Performance Evaluation for Tracking and Surveillance, Rio de Janeiro (2007)
15. Guadarrama, S., Dahl, R., Bieber, D., Norouzi, M., Shlens, J., Murphy, K.: Pixcolor: Pixel recursive colorization. arXiv preprint arXiv:1705.07208 (2017)
16. Guo, C.C., Chen, S.Z., Lai, J.H., Hu, X.J., Shi, S.C.: Multi-shot person re-identification with automatic ambiguity inference and removal. In: 2014 22nd International Conference on Pattern Recognition, pp. 3540–3545 (2014)
17. Herrmann, C., Müller, T., Willersinn, D., Beyerer, J.: Real-time person detection in low-resolution thermal infrared imagery with MSER and CNNs. In: Huckridge, D.A., Ebert, R., Lee, S.T. (eds.) SPIE Security + Defence, p. 99870I–8. SPIE, October 2016
18. Hirzer, M., Beleznai, C., Roth, P.M., Bischof, H.: Person re-identification by descriptive and discriminative classification. In: Heyden, A., Kahl, F. (eds.) SCIA 2011. LNCS, vol. 6688, pp. 91–102. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-21227-7_9
19. Hwang, S., Park, J., Kim, N., Choi, Y., So Kweon, I.: Multispectral pedestrian detection: benchmark dataset and baseline. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2015
20. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5967–5976. IEEE (2017)
21. John, V., Tsuchizawa, S., Liu, Z., Mita, S.: Fusion of thermal and visible cameras for the application of pedestrian detection. Sig. Image Video Process. **11**(3), 517–524 (2016)
22. Jojic, N., Perina, A., Cristani, M., Murino, V., Frey, B.: Stel component analysis: modeling spatial correlations in image class structure. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, pp. 2044–2051. IEEE (2009)
23. Kniaz, V.V., Gorbatsevich, V.S., Mizginov, V.A.: Thermalnet: a deep convolutional network for synthetic thermal image generation. In: ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. XLII-2/W4, pp. 41–45 (2017). https://doi.org/10.5194/isprs-archives-XLII-2-W4-41-2017
24. Kniaz, V.V., Mizginov, V.A.: Thermal texture generation and 3D model reconstruction using SFM and GAN. In: ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. XLII-2, pp. 519–524 (2018). https://doi.org/10.5194/isprs-archives-XLII-2-519-2018
25. Knyaz, V.A., et al.: Deep learning of convolutional auto-encoder for image matching and 3D object reconstruction in the infrared range. In: The IEEE International Conference on Computer Vision (ICCV) Workshops, October 2017
26. Larsen, A.B.L., Sønderby, S.K., Larochelle, H., Winther, O.: Autoencoding beyond pixels using a learned similarity metric. In: Balcan, M.F., Weinberger, K.Q. (eds.) Proceedings of the 33rd International Conference on Machine Learning. Proceedings of Machine Learning Research, PMLR, New York, vol. 48, pp. 1558–1566, 20–22 June 2016. http://proceedings.mlr.press/v48/larsen16.html

27. Li, C., Wand, M.: Precomputed Real-Time Texture Synthesis with Markovian Generative Adversarial Networks. arXiv.org, April 2016
28. Li, W., Zhao, R., Wang, X.: Human reidentification with transferred metric learning. In: Lee, K.M., Matsushita, Y., Rehg, J.M., Hu, Z. (eds.) ACCV 2012. LNCS, vol. 7724, pp. 31–44. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-37331-2_3
29. Li, W., Zhao, R., Xiao, T., Wang, X.: DeepReID: deep filter pairing neural network for person re-identification. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition, pp. 152–159. Chinese University of Hong Kong, Hong Kong. IEEE, January 2014
30. Limmer, M., Lensch, H.P.: Infrared colorization using deep convolutional neural networks. In: 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 61–68. IEEE (2016)
31. Matas, J., Chum, O., Urban, M., Pajdla, T.: Robust wide baseline stereo from maximally stable extremal regions. In: Proceedings of the British Machine Vision Conference, pp. 36.1–36.10. British Machine Vision Association (2002)
32. Morerio, P., Cavazza, J., Murino, V.: Minimal-entropy correlation alignment for unsupervised deep domain adaptation. arXiv preprint arXiv:1711.10288 (2017)
33. Nguyen, D., Hong, H., Kim, K., Park, K.: Person recognition system based on a combination of body images from visible light and thermal cameras. Sensors **17**(3), 605–29 (2017)
34. Nguyen, D., Kim, K., Hong, H., Koo, J., Kim, M., Park, K.: Gender recognition from human-body images using visible-light and thermal camera videos based on a convolutional neural network for image feature extraction. Sensors **17**(3), 637–22 (2017)
35. Nguyen, D., Park, K.: Body-based gender recognition using images from visible and thermal cameras. Sensors **16**(2), 156–21 (2016)
36. Nguyen, D., Park, K.: Enhanced gender recognition system using an improved histogram of oriented gradient (HOG) feature from quality assessment of visible light and thermal images of the human body. Sensors **16**(7), 1134–25 (2016)
37. Nguyen, D.T., Hong, H.G., Kim, K.W., Park, K.R.: Person recognition system based on a combination of body images from visible light and thermal cameras. Sensors **17**(3), 605 (2017)
38. Paszke, A., et al.: Automatic differentiation in pytorch (2017)
39. Paul, A., Vogt, K., Rottensteiner, F., Ostermann, J., Heipke, C.: A comparison of two strategies for avoiding negative transfer in domain adaptation based on logistic regression. In: International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives, pp. 845–852. Gottfried Wilhelm Leibniz Universitat, Hannover, Germany, May 2018
40. Prosser, B., Gong, S., Xiang, T.: Multi-camera matching using bi-directional cumulative brightness transfer functions. In: Proceedings of the British Machine Vision Conference, BMVC 2008, pp. 64.1–64.10. Queen Mary, University of London, London, United Kingdom, British Machine Vision Association, January 2008
41. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
42. San-Biagio, M., Ulas, A., Crocco, M., Cristani, M., Castellani, U., Murino, V.: A multiple kernel learning approach to multi-modal pedestrian classification. In: 2012 21st International Conference on Pattern Recognition (ICPR), pp. 2412–2415. IEEE (2012)

43. St-Laurent, L., Maldague, X., Prévost, D.: Combination of colour and thermal sensors for enhanced object detection. In: 2007 10th International Conference on Information Fusion, pp. 1–8. IEEE (2007)
44. Ulyanov, D., Lebedev, V., Vedaldi, A., Lempitsky, V.S.: Texture networks - feed-forward synthesis of textures and stylized images. CoRR abs/1501.02565 1603, arXiv:1603.03417 (2016)
45. Méndez, H., Martín, C.S., Kittler, J., Plasencia, Y., García-Reyes, E.: Face recognition with LWIR imagery using local binary patterns. In: Tistarelli, M., Nixon, M.S. (eds.) ICB 2009. LNCS, vol. 5558, pp. 327–336. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-01793-3_34
46. Vogt, K., Paul, A., Ostermann, J., Rottensteiner, F., Heipke, C.: Unsupervised source selection for domain adaptation. Photogrammetric Eng. Remote Sens. **84**, 249–261 (2018)
47. Wu, A., Zheng, W.S., Yu, H.X., Gong, S., Lai, J.: RGB-infrared cross-modality person re-identification. In: The IEEE International Conference on Computer Vision (ICCV), October 2017
48. Wu, J., Wang, Y., Xue, T., Sun, X., Freeman, W.T., Tenenbaum, J.B.: MarrNet: 3D Shape Reconstruction via 2.5D Sketches. arXiv.org, November 2017
49. Xie, Z., Jiang, P., Zhang, S.: Fusion of LBP and HOG using multiple kernel learning for infrared face recognition. In: ICIS (2017)
50. Xu, D., Ouyang, W., Ricci, E., Wang, X., Sebe, N.: Learning cross-modal deep representations for robust pedestrian detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4236–4244. IEEE, April 2017
51. Ye, M., Lan, X., Li, J., Yuen, P.C.: Hierarchical discriminative learning for visible thermal person re-identification. In: AAAI (2018)
52. Ye, M., Wang, Z., Lan, X., Yuen, P.C.: Visible thermal person re-identification via dual-constrained top-ranking. In: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, pp. 1092–1099. International Joint Conferences on Artificial Intelligence Organization, California (2018)
53. Yilmaz, A., Shafique, K., Shah, M.: Tracking in airborne forward looking infrared imagery. Image Vis. Comput. **21**, 623–635 (2002)
54. Zhang, H., Patel, V.M., Riggan, B.S., Hu, S.: Generative adversarial network-based synthesis of visible faces from polarimetrie thermal faces. In: 2017 IEEE International Joint Conference on Biometrics (IJCB), pp. 100–107. IEEE (2017)
55. Zhang, M.M., Choi, J., Daniilidis, K., Wolf, M.T., Kanan, C.: VAIS - a dataset for recognizing maritime imagery in the visible and infrared spectrums. In: CVPR Workshops, pp. 10–16 (2015)
56. Zhang, R., Isola, P., Efros, A.A.: Colorful image colorization. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9907, pp. 649–666. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46487-9_40
57. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018
58. Zhang, T., Wiliem, A., Yang, S., Lovell, B.C.: TV-GAN: Generative Adversarial Network Based Thermal to Visible Face Recognition, December 2017
59. Zheng, L., et al.: MARS: a video benchmark for large-scale person re-identification. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9910, pp. 868–884. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46466-4_52

60. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: a benchmark. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1116–1124. Tsinghua University, Beijing, China. IEEE, February 2015
61. Zheng, W.S., Gong, S., Xiang, T.: Associating groups of people. In: British Machine Vision Conference (2009)
62. Zhu, J.Y., et al.: Toward multimodal image-to-image translation. In: Advances in Neural Information Processing Systems, pp. 466–477. University of California, Berkeley, United States, January 2017
63. Zhu, J.Y., et al.: Toward multimodal image-to-image translation. In: Guyon, I., et al. (eds.) Advances in Neural Information Processing Systems, vol. 30, pp. 465–476. Curran Associates, Inc. (2017). http://papers.nips.cc/paper/6650-toward-multimodal-image-to-image-translation.pdf