Learning Modal-Invariant and Temporal-Memory for Video-based Visible-Infrared Person Re-Identification

Xinyu Lin[1], Jinxing Li[1] ✉, Zeyu Ma[1], Huafeng Li[2], Shuang Li[2],Kaixiong Xu[2], Guangming Lu[1], David Zhang[3]
[1]Harbin Institute of Technology, Shenzhen, [2]Kunming University of Science and Technology,
[3]The Chinese University of HongKong, Shenzhen.
{linxinyu0327, lijinxing158}@gmail.com, zeyu.ma@stu.hit.edu.cn, hfchina99@163.com,
{shuangli936, xukaixiong99}@gmail.com, luguangm@hit.edu.cn, davidzhang@cuhk.edu.cn

Abstract

Thanks for the cross-modal retrieval techniques, visible-infrared (RGB-IR) person re-identification (Re-ID) is achieved by projecting them into a common space, allowing person Re-ID in 24-hour surveillance systems. However, with respect to the probe-to-gallery, almost all existing RGB-IR based cross-modal person Re-ID methods focus on image-to-image matching, while the video-to-video matching which contains much richer spatial- and temporal-information remains under-explored. In this paper, we primarily study the video-based cross-modal person Re-ID method. To achieve this task, a video-based RGB-IR dataset is constructed, in which 927 valid identities with 463,259 frames and 21,863 tracklets captured by 12 RGB/IR cameras are collected. Based on our constructed dataset, we prove that with the increase of frames in a tracklet, the performance does meet more enhancement, demonstrating the significance of video-to-video matching in RGB-IR person Re-ID. Additionally, a novel method is further proposed, which not only projects two modalities to a modal-invariant subspace, but also extracts the temporal-memory for motion-invariant. Thanks to these two strategies, much better results are achieved on our video-based cross-modal person Re-ID. The code and dataset are released at: https://github.com/VCM-project233/MITML.

1. Introduction

Person re-identification (Re-ID) [17, 25, 43, 54] focuses on matching probe pedestrian images with the gallery sets. Due to multiple views which are non-overlapped, there are significant changes in human body postures, illumination and backgrounds, leading a large challenge to Re-ID. Thanks to the rapid development of deep learning, various deep end-to-end approaches [3, 22–24] have been studied, greatly enhancing the Re-ID performance.

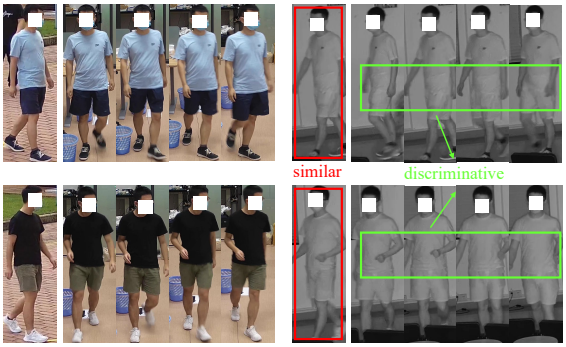

Figure 1. Advantages of video-based cross-modal person Re-ID. If two persons enjoy similar appearances, video data can also provide discriminative temporal-information that image data is unavailable. Specifically, the person wearing the black T-shirt is quite similar to the person wearing the blue T-shirt under the IR camera (shown in the red box), while their specific arm postures in the motions give the discriminative features (shown in the green box).

Despite the achievement of aforementioned methods, most of them are heavily dependent on the RGB images, so that the lighting for the cameras is essential. However, this constraint is too strict, especially at night, making the collected RGB data uninformative and failing to achieve person Re-ID. Fortunately, most surveillance cameras can automatically switch from RGB to the infrared (IR) mode if the lighting is unavailable. In contrast to RGB images, IR images are capable of preserving the information under invisible lighting and showing pedestrians clearly. Thus, in order to achieve person Re-ID in 24-hour surveillance systems, the RGB-IR based visible-infrared (cross-modal) person Re-ID [6, 25, 39] provides a promising strategy. For instance, Wu *et al.* [39] first collected an RGB-IR dataset and proved the feasibility for these two modalities matching. Inspired by this work, various cross-modal Re-ID works were

✉ Jinxing Li is the Corresponding Author.

then studied.

Although the cross-modal person Re-ID methods fill the gap between RGB images and IR images, they are only single-image based tasks. In the data collection, pedestrians originally appear in the video databases, containing multiple frames in each tracklet. Intuitively, the video-based data contains much richer visual information than a single image [53]. In some specific cases, it is indeed difficult to identify two persons with similar appearances if only a single image is given. This case is more difficult for the infrared modality, and even the human beings cannot guarantee the correctness. In contrast to still images, the video is an image sequence containing the spatial and temporal-information, so that the beneficial motion information can be exploited for the discriminative identification. For instance, as displayed in Fig. 1, two images captured from two persons enjoy similarity under the IR camera. However, the person wearing the black T-shirt has the specific arm posture in the motion, compared with the person wearing the blue T-shirt. Thanks to such motion characteristics, more discriminative information is provided for us to achieve a more robust and accurate identification model. Thus, it is quite significant to replace the still-images with videos in cross-modal person Re-ID.

To address this problem, in this paper, the video-based cross-modal Re-ID is studied. In comparison to image-based cross-modal Re-ID, the video-based cross-modal Re-ID further aims to exploit the temporal-information for robust feature extraction. In contrast to the existing video-based RGB Re-ID methods, our focused work additionally extracts the consistency between RGB and IR modalities.

In order to achieve the video-based cross-modal Re-ID, an associated database is inevitable. Although Wu et al. [39] has presented an RGB-IR dataset, it only focuses on the image-based retrieval, being far away from our video-based requirement. To substantiate our task, we primarily construct a video-based RGB-IR database named HITSZ Video Cross-Modal (HITSZ-VCM) Re-ID dataset. The comparison between our collected dataset and existing Re-ID datasets is listed in Tab. 1. Different from SYSU-MM01 [39] which only collected the RGB images and IR images via 4 RGB cameras and 2 IR cameras, we set 12 cameras to capture both RGB and IR videos and much more valid identities are collected. Totally, 927 valid identities including 11,785 / 10,078 tracklets and 251,452 / 211,807 images with or free from masks for RGB and IR modalities are obtained, respectively.

For the video-based cross-modal Re-ID, the spatial- and temporal-information among each tracklet do contribute to the performance improvement. In this paper, a baseline method is first applied to our constructed dataset, demonstrating the significance of video-based cross-modal Re-ID. Specifically, we follow Ye *et al.*'s [49] baseline on image-based cross-modal Re-ID and add a module to utilize the temporal-information. Additionally, we also propose a novel method named Modal-Invariant and Temporal-Memory Learning (MITML). Two modalities are transformed to get modal-invariant but id-related features through an adversarial strategy, so that the gap between RGB and IR modalities is relieved. Referring to the motion information in a tracklet, we also propose a temporal memory refinement module to extract the temporal-information. Thanks to these two strategies, the Re-ID performance on our dataset is further improved.

Overall, the main contributions of this paper are:

- We construct a video-based RGB-IR database, allowing the study on video-based cross-modal person Re-ID. **Different from existing Re-ID works, to the best of our knowledge, this is the first work which jointly takes cross modalities and videos into account, defining a challenging task.**
- We introduce a baseline to prove the significance of video-based cross-modal person Re-ID. In detail, by embedding a temporal-information exploitation module, the cross-modal person Re-ID performance meets a continuous increase when the number of images in a tracklet rises.
- A novel method named Modal-Invariant and Temporal-Memory Learning (MITML) is additionally proposed by more efficiently removing modal-variance and exploiting motion information. Experimental results substantiate the superiority of our proposed method.

2. Related Works

2.1. Visible-infrared Person Re-ID

Visible-infrared person Re-ID handles person retrieval between different modalities, which is implemented by setting RGB cameras and IR cameras. Considering the poor illumination condition in some cases, especially at night, this cross-modal task does enjoy practical significance. Thanks to the image-based RGB-IR person Re-ID dataset constructed by Wu *et al.* [39], many cross-modal Re-ID technologies have been studied, most of which are based on metric learning [9, 13, 18, 26, 45–47, 51], feature learning [31, 44–47, 49, 51], and adversarial learning [4, 4, 6, 28, 28, 33, 36, 38], etc.

As for metric learning and feature learning, Ye *et al.* [46] extracted multi-modal shareable features in the feature learning stage, after which heterogenous features are projected into a common space and measured by the metric learning. Hao *et al.* [13] maps the extracted features onto a hypersphere manifold, in which differences be-

tween two samples are calculated based on angles. Different from angle measurement in [13], Feng *et al.* [9] utilized the Euclidean constrain to shrink the cross-modal gap. In [51], inter-modality and intra-modality variations are firstly taken into account, and the discriminative features are then learned through a top-ranking loss. Ye *et al.* [47] then extended [51] by introducing a bi-directional center-constrained top-ranking loss, further improving the image based person Re-ID performance.

Additionally, the generative adversarial network (GAN) [11] has also been widely applied to cross-modal Re-ID tasks. Dai *et al.* [6] embedded a discriminator into the network, enforcing the features from two modalities to be unclassified in an adversarial way. In [36], by introducing the CycleGAN [55], an RGB image is transformed to an IR version, while its id-information is preserved. So is the IR image. Furthermore, some researchers [4, 28] also achieved feature learning in an adversarial and disentanglement learning way. Particularly, Choi *et al.* [4] disentangled id-discriminative features and id-excluded features from cross-modal images, which are then combined to generate modal-different but id-consistent images. However, this strategy encounters a large computational complexity and some generated images with poor quality do give an inferior influence on the performance.

Apart from aforementioned methods, Ye *et al.* [48] additionally focused on the image properties of RGB/ IR images. For instance, a novel joint learning strategy by channel augmentation and simulating random occlusions is proposed. Moreover, image alignment [31] and pattern alignment [40] are also exploited to alleviate the discrepancies.

2.2. Video-based Person Re-ID

Different from image-based person Re-ID, video-based person Re-ID represents a person by a sequence of images, providing temporal-information and a richer appearance [50]. Generally, existing methods mainly adopt RNN [27, 29, 42, 52], temporal pooling (average or weighted) [5, 10, 41, 42], optical flow [2, 29, 52], and 3D convolution [12, 21, 27], etc. For instance, Xu *et al.* [42] took original person images and corresponding optical flow as the network inputs, so that the motion consistency for a person in different periods is ensured. Then, features extracted from the CNN-RNN module are utilized to compute attention vectors, selecting informative frames over the sequence. A novel Spatial and Temporal Memory Networks (STMN) [8] is proposed, in which features for spatial distractors that frequently emerge across video frames, as well as attentions optimized for typical temporal patterns, are both stored. Furthermore, Aich *et al.* [1] designed a flexible feature processing module which can be used in any 3D convolutional block for the Re-ID task. Thanks to this module, complementary person-specific appearance and motion information are well captured. Besides, 3D graph convolution is also introduced for video-based Re-ID. Liu *et al.* [27] employed context-reinforced topology to build a graph, which successfully encodes contextual information and physical information of the human body. By applying the 3D graph convolutional layers to it, spatial-temporal dependencies and structural information are efficiently captured.

Despite the fact that a number of works have been done for person ID, they are only either image-based cross-modal Re-ID or video-based RGB Re-ID. Our HITSZ-VCM is the first dataset combining cross modalities and video data, allowing the study on video-based cross-modal person Re-ID. It not only achieves the 24-hour surveillance, but also gets comprehensive information and obtains much higher Re-ID accuracy.

3. Dataset

3.1. Dataset Description

In this paper, we build the HITSZ-VCM (HITSZ Video Cross-Modal) Re-ID dataset for the video-based cross-modal person Re-ID task. To our best knowledge, this is the first cross-modal Re-ID dataset based on videos. The HITSZ-VCM dataset contains a large amount of images/frames captured by 12 HD cameras with 3840×2160 resolution. Thanks to the modern monitoring technology, all the cameras can automatically shoot both RGB and IR images according to the lighting conditions. Thus, each person is captured by both RGB and IR cameras. Note that all tracklets are processed by an automatic object tracking system, and we then finetune inaccurate annotations manually.

In detail, our HITSZ-VCM dataset contains 927 valid identities. The cameras shoot 25 frames per second and we extract the first frame out of every 5 frames to build the final dataset. According to this setting, every 24 consecutive images are regarded as a tracklet for a person during the same period, and the last frames whose number may be less than 24 form the last tracklet. Totally, there are 251,452 RGB images and 211,807 IR images, which can be divided into 11,785 and 10,078 tracklets, respectively. Of course, the number of frames in a tracklet can also be dynamically set, which is more flexible than many existing video-based datasets. More specifically, 12 cameras are used for our video collection. Generally, most of identities are captured by 3 RGB cameras and 3 IR cameras, and these cameras are non-overlapped.

Our HITSZ-VCM dataset also covers a series of diverse scenarios. Firstly, 7 outdoor, 3 indoor and 2 passages scenes are included. In detail, some common venues like the office, cafe, passageway, playground, and garden are all considered. Besides, each person is captured from multiple angles under each camera, constructing a richer appearance set.

Table 1. Comparison between HITSZ-VCM with some typical Re-ID datasets.

Dataset	Type	#Identites	#RGB cam.	#IR cam.	#Images & BBoxes	#Tracklets	Evaluation
iLIDS-VID [37]	Video	300	2	0	42,495	600	CMC
MARs [53]	Video	1,261	6	0	1,067,516	20,715	CMC + mAP
Duke-Video [41]	Video	1,812	8	0	815,420	4,832	CMC + mAP
LS-VID [20]	Video	3,772	15	0	2,982,685	14,943	CMC + mAP
RegDB [30]	Image	412	1	1	8,240	-	CMC + mAP
SYSU-MM01 [39]	Image	491	4	2	303,420	-	CMC + mAP
HITSZ-VCM	Video	927	12	12	463,259	21,863	CMC + mAP

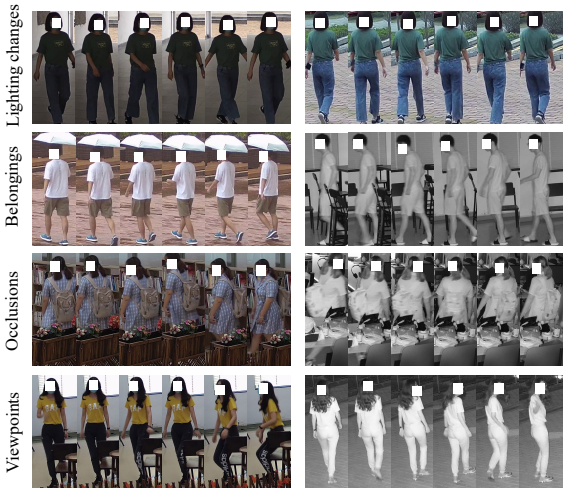

Figure 2. Some challenging tracklets in our dataset, including lighting changes, belonging changes, occlusions and viewpoint changes.

Furthermore, some challenging scenarios such as lighting changes (for RGB images), belonging changes, occlusion and viewpoint changes are collected, as displayed in Fig. 2.

Tab. 1 tabulates the comparison between HITSZ-VCM and existing related Re-ID datasets. As we can see, although MARs [53], Duke-Video [41], and LS-VID [20] also enjoy a large number of valid identities, they fail to cover IR images or videos, being incapable for the 24-hour surveillance. In contrast to RegDB [30] and SYSU-MM01 [39], our constructed HITSZ-VCM dataset extends the image-based version to the video-based one, which provides more abundant and valuable information for person Re-ID. Furthermore, there are much more identities captured from more diverse scenarios in our dataset, greatly contributing to the training of the deep network.

In conclusion, HITSZ-VCM enjoys the following characteristics: (1) Constructing the first video-based cross-modal dataset for person Re-ID. (2) Collecting much more valid identities under diverse scenes. (3) Covering challenging but practical cases.

3.2. Evaluation Protocol

Here we conduct cross-camera and cross-modal retrieval like existing works [20, 30, 39, 53]. In other words, query and gallery are captured by different cameras and modalities. Meanwhile, with respect to the 'probe to gallery' pattern, video-to-video matching is adopted to keep consistent with training data. Regularly, we utilize two retrieval modes for HITSZ-VCM: 'infrared to visible' and 'visible to infrared', to achieve a more comprehensive evaluation. Additionally, we take all the tracklets in one modality as the query set and those from the other modality as the gallery set. Totally, in the 'infared to visible' retrieval, there respectively exist 5,159 and 5,643 tracklets in the query set and the gallery set. Vice versa in the 'visible-to-infrared' mode. Note that in our implementation, we discard some too short tracklets (less than 12 images).

To quantitatively evaluate the performance on our proposed dataset, the Cumulative Matching Characteristic curve (CMC) and mean Average Precision(mAP) are adopted as the evaluation metrics. Being similar to many methods, we compute the distance scores of all the query features and gallery features to do the ranking work. For testing, cosine similarity is used as the distance measurement.

4. Baseline

We follow the baseline proposed in [49]. A two-stream network, with ResNet50 [14] utilized as the backbone [36, 38, 49], is employed to handle the heterogeneous data belonging to different modalities. Specifically, the first convolutional blocks in two branches enjoy different weights, so that modal-specific features for RGB and IR sequences are learned, respectively. Differently, in the remaining four blocks, the weights are shared to extracted modal-invariant features for these two modalities. Since the inputs of the network are multiple images, an average pooling layer is utilized to fuse the frame-level features obtained from the backbone. Thus, the sequence-level feature of each tracklet is finally obtained. By following [49], the identity loss is introduced to guide the intra-modal Re-ID task, while the

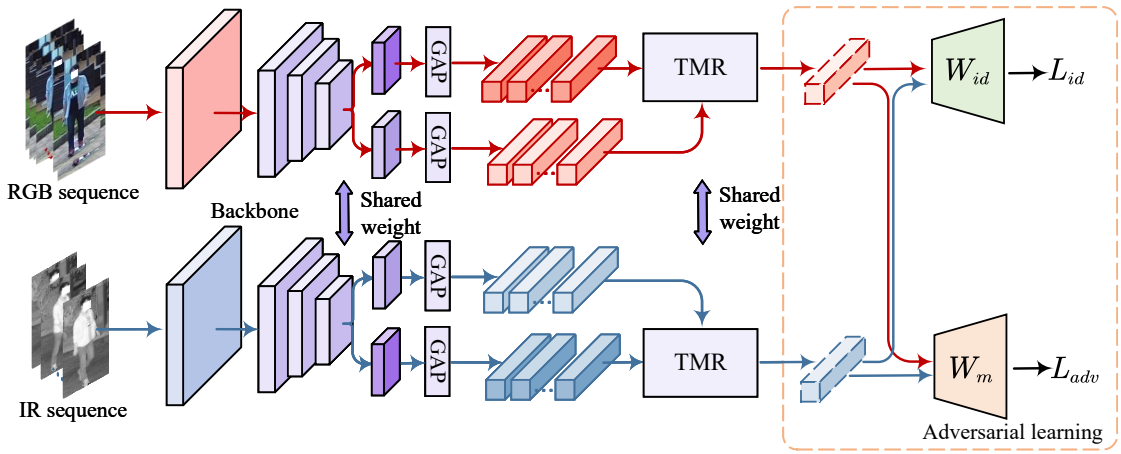

Figure 3. The framework of our proposed method. RGB/IR image sequences are regarded as corss-modal inputs. The Temporal Memory Refinement (TMR) module aggregates frame-level features into sequence-level features. W_{id} and W_m lead the classification of identities and modalities. By adopting an adversarial learning strategy, modal-related information is removed from the cross-modal data, and only id-related features are preserved.

triplet loss is exploited to handle the cross-modal Re-ID task. Therefore, the objective function $\mathcal{L}^{base}$ can be formulated as follows:

$$\mathcal{L}^{base} = \mathcal{L}_{id}^{base} + \mathcal{L}_{tri}^{base} \tag{1}$$

where $\mathcal{L}_{id}^{base}$ and $\mathcal{L}_{tri}^{base}$ denote the identity loss and the triplet loss, respectively.

5. Proposed Method

Based on the baseline, a novel method is further proposed to more efficiently learn the modal-invariant and time-memory features for RGB and IR modalities. The framework of our model is illustrated in Fig. 3. By modifying the last convolution block in the backbone to two-branch convolution blocks (shared structures but different weights), two sets of feature maps from a sequence are obtained and then forwarded into a Temporal Memory Refinement (TMR) module, so that the temporal-information is extracted to meet the motion consistency for an identity. Furthermore, to fill the gap between two modalities, two classifiers are introduced, through which the modality-related features are removed while the id-related features are enhanced, greatly contributing to the cross-modal retrieval.

5.1. Temporal Memory Refinement

Here we respectively denote an RGB sequence and an IR sequence as $\mathbf{V} = \left\{\mathbf{V}^t \middle| \mathbf{V}^t \in \mathbb{R}^{H\times W}\right\}_{t=1}^{T}$ and $\mathbf{I} = \left\{\mathbf{I}^t \middle| \mathbf{I}^t \in \mathbb{R}^{H\times W}\right\}_{t=1}^{T}$, where H and W denote the height and weight of the images, t means the t-th frame of this sequence, and T is the total number of images in a tracklet. Correspondingly, the ID labels are denoted as p_v and p_i, while m_v and m_i denote the modal labels.

To transform frame-level features into a sequence-level feature and effectively capture temporal contexts among multiple frames, inspired by [8], we propose a Temporal Memory Refinement (TMR) module. The structure of TMR is shown in Fig. 4. LSTM [15] layers and the SE attention [16] jointly facilitate the exploitation of temporal information, refining the features to enjoy more discriminative information.

Take RGB data $\mathbf{V}$ as an example and the five convolution blocks in our backbone are denoted as E_{res}. We denote the two sets of frame-level features from E_{res} as $\mathbf{f}_{v1} = \{\mathbf{f}_{v1}^t\}_{t=1}^{T}$ and $\mathbf{f}_{v2} = \{\mathbf{f}_{v2}^t\}_{t=1}^{T}$, which are utilized for the attention weights generation and the frame-level features aggregation, respectively. Features $\mathbf{f}_{v1}$ are first forwarded into two LSTM layers $LSTM^2$, so that the temporal context of this tracklet is obtained. By applying a full-connected layer FC^t to its associated output from $LSTM^2$ and adding $\mathbf{f}_{v1}^t$, an attention $\mathbf{a}^t$ is obtained by following the SE attention module.

$$\mathbf{a}^t = SE^t\left(\left(FC^t\left(LSTM^2(\mathbf{f}_{v1})\right) + \mathbf{f}_{v1}^t\right)/2\right), \tag{2}$$

where SE^t means the t-th SE attention module. Note that $\mathbf{a}^t$ denotes the temporal-memory which gives the importance of the t-th frame in a tracklet. In other words, $\mathbf{a}^t$ plays as the attention weight of the t-th frame-level feature.

Based on the aforementioned analysis, the person representation of the t-th frame could be refined and the average pooling processing is then utilized to aggregate the frame-level features into a sequence-level one:

$$\mathbf{F}_v = \sum_{t=1}^{T}\left(\mathbf{a}^t \odot \mathbf{f}_{v2}^t + \mathbf{f}_{v2}^t\right)/T \tag{3}$$

Similar processing is conducted for IR data with the shared weights, through which its sequence-level $\mathbf{F}_i$ is obtained.

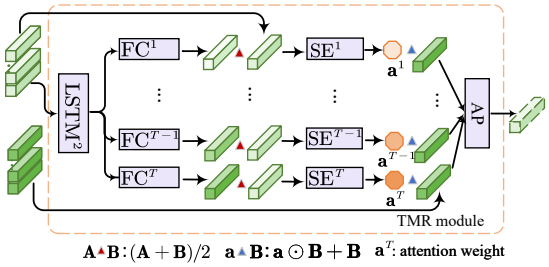

Figure 4. The insight of the TMR module. Multiple images from a tracklet are refined and the frame-level features are aggregated into a sequence-level one.

Overall, the temporal information is captured by the TMR module, so that the person representations are refined within a single modality. In our training phase, this module is optimized simultaneously with E_{res}, which is regarded as a supplementary for E_{res}.

5.2. Modal-Invariant Learning

After exploiting TMR where the intra-modal features are refined, we then introduce a modal-invariant adversarial learning to remove the gap from two modalities. Referring to the adversarial strategy, AlignGAN proposed in [36] transforms the RGB/IR image to the IR/RGB version in the pixel-level by confusing a modal-discriminator. However, this strategy is constrained on the image generation, which not only increases the computational complexity, but also is quite sensitive to the quality of the generated image. By contrast, inspired by [19], here we achieve the adversarial learning only based on the feature-level, more efficiently getting the modal-invariant features.

According to Eq.(3), the sequence-level features for RGB and IR tracklets are $\mathbf{F}_v$ and $\mathbf{F}_i$, respectively. Theoretically, if $\mathbf{F}_v$ and $\mathbf{F}_i$ do enjoy the id-related information but without modal-related information, they cannot be classified to m_v or m_i. To achieve this task, a classifier W_m is introduced whose output is a 3×1 vector. Particularly, this output denotes the probabilities of a tracklet belonging to the RGB modality, IR modality, or neither of them. The objective function is formulated as:

$$\mathcal{L}_{adv1}(E) = CE(W_m(\mathbf{F}_v), m_3) + CE(W_m(\mathbf{F}_i), m_3) \quad (4)$$

where E is the combination of the backbone E_{res} and TMR, $CE(\cdot)$ denotes the cross-entropy loss, and m_3 means the third category which is neither belonging to the RGB modality m_v nor the IR modality m_i. To encourage $\mathbf{F}_v$ and $\mathbf{F}_i$ to enjoy the discriminative information on identities, we also apply the id-related cross entropy loss and triplet loss to them via another classifier W_{id}. Thus, the id-related but modal-invariant function can be represented as:

$$\mathcal{L}_{id}(E, W_{id}) = \mathcal{L}_{adv1} + \mathcal{L}_{id}^{ce} + \mathcal{L}_{id}^{tri} \quad (5)$$

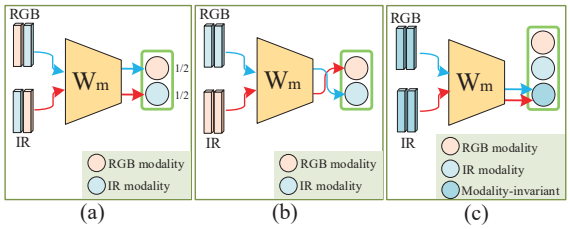

Figure 5. Different strategies in adversarial learning. (a) [34]; (b) [6]; (c) Ours.

where $\mathcal{L}_{id}^{ce}$ and $\mathcal{L}_{id}^{tri}$ are the id-related cross-entropy loss and triplet loss, respectively.

Of course, the classification capability of W_m plays a key role for the modality-invariant feature learning. Here, being similar to existing GAN based methods, the adversarial learning strategy is adopted to additionally update W_m, shown as follows:

$$\mathcal{L}_{adv2}(W_m) = CE(W_m(\mathbf{F}_v), m_v) + CE(W_m(\mathbf{F}_i), m_i) \quad (6)$$

In the optimization, Eq.(5) and Eq.(6) are optimized in an alternative way, so that W_m enjoying the more strong capacity adversarially contributes to the modal-invariant feature learning. Note that, it is true [34] and [6] also learn the id-related features via the adversarial learning in the feature level. However, they are different from each other. As displayed in Fig.5(a), [34] enforces the probability belonging to each category to be the same, while a feature which contains both RGB and IR information can also have the same classification result. Obviously, in this case, we are unsure that whether these two modalities are aligned. Referring to Fig.5(b), RGB and IR features are inversely classified to IR and RGB modalities. A limitation is that these two inputs are transformed to only gain each other modal-specific information, while the modal-gap is not measured. By contrast, our used strategy directly classifies different features into an additional class, guaranteeing that they do fall in a same latent space or domain.

6. Experiment

6.1. Experimental Setting

Dataset. Here we divide our dataset into two sets for training and testing. The training set contains 500 identities with 232,496 images and 11,061 tracklets, while the testing set contains 427 identities with 230,763 images and 10,802 tracklets. In the training phase, all the images are resized to a size of 288×144. Being similar to many existing methods, random cropping with zero-padding and horizontal flipping are also used for the data augmentation.

Experimental Implementation. We implement our model by PyTorch [32] and train it on a NVIDIA TESLA A100

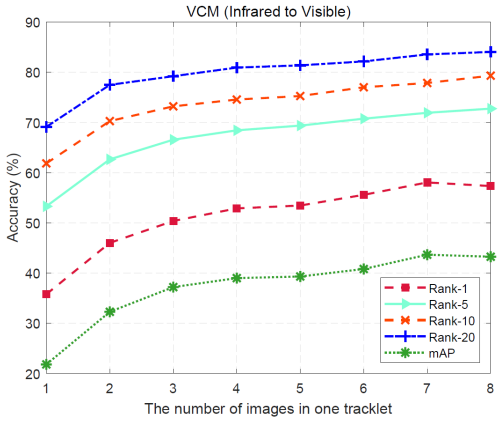

Figure 6. Evaluation of our baseline on different settings, where n denotes the number of images in one tracklet.

Table 2. Effectiveness of the TMR module and the adversarial learning module. Note that M denotes the modal-invariant learning, T denotes the temporal memory refinement(TMR), and Full method^S denotes full method with shuffled frames in TMR.

	Infrared to Visible				
Strategy	R1	R5	R10	R20	mAP
Baseline	55.58	70.75	77.01	82.16	40.80
Baseline + M	59.73	74.50	80.06	84.69	42.80
Baseline + T	58.44	72.32	78.51	83.51	43.87
Full method^S	60.82	74.54	78.69	83.25	43.94
Full method	**63.74**	**76.88**	**81.72**	**86.28**	**45.31**

with cuda version 11.2. The ResNet50 [14] pretrained on the ImageNet [7] is exploited our baseline and backbone. For the encoder E and id-classifier W_{id}, they are optimized via the optimizer SGD with the weight decay of 5×10^{-4} and the momentum of 0.9. We adopt a learning rate warmup strategy for E and W_{id} and its initial value is set to 0.1. After 35 epochs and 80 epochs, the learning rate is reduced to 0.01 and 0.001, respectively. Note that the learning rate of the first unshared convolution blocks is always the one tenth of that of the remaining modules. In terms of the modal-classifier W_m, the SGD optimizer works with the 0.01 learning rate, the 5×10^{-4} weight decay and 0.9 momentum. We set the maximum number of epochs to 200. Besides, the batch size is set to 16 with 8 different identities and 2 tracklets for each identity.

Furthermore, for each tracklet with 24 continuous images, n (it can be dynamically selected) images are selected for training. Specifically, 24 images are divided into n parts with $24/n$ images, in which 1 of $24/n$ images is randomly selected to form the training data.

6.2. Ablation Study

In this subsection, we experimentally analyze the significance of our HITSZ-VCM dataset, as well as the strategies in the proposed method MITML.

Figure 7. The visualization results on different settings, where n denotes the number of images in one tracklets. Obviously, image-based methods ($n = 1$) cannot report satisfying results when two identities enjoy similar appearances, while video-based methods show noticeable performance with the enhancement of temporal information.

Table 3. Comparisons of our modal-invariant learning with different adversarial strategies shown in Fig. 5. For a fair comparison, we only replace our modal-invariant learning with other strategies in our network.

	Infrared to Visible				
Strategy	R1	R5	R10	R20	mAP
cmGAN [6]	57.96	72.58	78.32	83.42	43.14
UCDA [34]	59.51	73.34	79.14	84.18	45.06
Our method	**63.74**	**76.88**	**81.72**	**86.28**	**45.31**

Significance of video-based cross-modal datasets. Compared with image data, video data provide more abundant information for the person Re-ID task. To verify the above statement, we conduct the experiments by changing the number of images in one tracklet on our baseline, as displayed in Fig. 6. It can be seen that with the increase of images in a tracklet, the Re-ID performance meet a continuous rise, demonstrating the significance of our constructed dataset. Specifically, when only one image is exploited, like that in existing image-based cross-modal Re-ID dataset, the mAP is only 23.09%, which is much inferior to that when six images are simultaneously used in a tracklet. Fig. 7 further illustrates the Top-10 visualization results of our proposed method MITML with different settings on datasets, also substantiating the necessity of our HITSZ-VCM dataset.

As shown in Fig. 6, the values of all metrics increase really slightly when n is relatively large, i.e., 7 and 8. To reduce the time costs in the training phase, we set n to 6 in the following experiments.

Table 4. Comparisons of our method with state-of-the-art cross-modal methods on our HITSZ-VCM dataset. CMC (%) and mAP (%) are reported.

Method	Venue	*Infrared to Visible*					*Visible to Infrared*				
		R1	R5	R10	R20	mAP	R1	R5	R10	R20	mAP
LbA [31]	ICCV'21	46.38	65.29	72.23	79.41	30.69	49.30	69.27	75.90	82.21	32.38
MPANet [40]	CVPR'21	46.51	63.07	70.51	77.77	35.26	50.32	67.31	73.56	79.66	37.80
DDAG [49]	ECCV'20	54.62	69.79	76.05	81.50	39.26	59.03	74.64	79.53	84.04	41.50
VSD [35]	CVPR'21	54.53	70.01	76.28	82.01	41.18	57.52	73.66	79.38	83.61	43.45
CAJL [48]	ICCV'21	56.59	73.49	79.52	84.05	41.49	60.13	74.62	79.86	84.53	42.81
Ours	-	**63.74**	**76.88**	**81.72**	**86.28**	**45.31**	**64.54**	**78.96**	**82.98**	**87.10**	**47.69**

Effectiveness of temporal-information exploitation. As shown in Tab. 2, we evaluate the performances of TMR in our approaches on our dataset. Compared with 'baseline', our TMR module ('Baseline + T' in Tab. 2) achieves a remarkable performance improvement on rank-1 and mAP, respectively. The main reason is that the temporal-information extracted by TMR facilitates to build an excellent appearance model and obtain person unique features which cannot be captured from image-based data. Also, we shuffle frames in a tracklet ('Full method^S'in Tab. 2), and the results are inferior, indicating the importance of temporal information.

Effectiveness of modal-invariant learning. The adversarial learning used in MITML successfully removes the modal-related information from different modalities but also preserves id-related features. As tabulated in Tab. 2, by adding this module into the baseline ('baseline + M'), there is indeed a performance enhancement, which also substantiates its effectiveness.

Furthermore, as shown in Tab. 3, we evaluate the other two adversarial learning strategies as discussed in Fig. 5. As we can see, the learning strategy in MITML is more effective than that in [34] and [6].

6.3. Comparison with State-of-the-art Methods

In this section, we further compare our proposed method with existing state-of-the-art visible-infrared cross-modal person Re-ID methods, including DDAG [49], LbA [31], MPANet [40] and VSD [35] and CAJL [48]. Note that, these comparison methods are primarily designed for image-based datasets. For a fair comparison, we conduct an average pooling layer for their generated frame-level features. For those networks whose backbones are ResNet50, including [31,35,40,49], we implement the average pooling after the backbone, which is similar to what we did in our baseline. Furthermore, we pretrain the model for CAJL [48] on AGW [50] before the channel augmented joint learning.

The CMC and mAP obtained by all these approaches on our dataset are listed in Tab. 4. Obviously, our method reports a noticeable improvement than these state-of-the-art imaged-based cross-modal approaches. Specifically, for the infrared to visible retrieval mode, Rank-1 and mAP increase by 7.15% and 3.82% respectively than the second best method CAJL. As for the visible to infrared search mode, Rank-1 and mAP also obtain a significant growth of 4.41% and 4.88%, indicating the effectiveness of our TMR module on the temporal information exploitation.

6.4. Limitation

Based on the above analysis, the importance of video-based cross-modal person Re-ID is proved, and our methods demonstrates a more remarkable improvement compared with existing methods. However, our methods requires a fixed number of images in one tracklet in the training and testing phases, which decreases the flexibility in realistic applications. In our future work, we will aim to design a novel network which can process tracklets with dynamic lengths.

7. Conclusion

Based on the observation that video data can provide temporal-information and allow us to build a richer appearance model for identification, we study a new task: video-based cross-modal person Re-ID. To achieve this goal, the first video-based cross-modal Re-ID dataset is constructed. There exist 927 valid identities with 251,452 RGB images of 11,785 tracklets and 211,807 IR images of 10,078 tracklets captured by 12 HD RGB/IR cameras, in which 500 identities for training and 427 identities for testing. Experimental results prove the significance of our constructed dataset. Additionally, a novel method: modal-invariant and temporal memory learning (MITML) is proposed for our HITSZ-VCM dataset. Specifically, an adversarial learning strategy contributes to extracting the high-quality modal-invariant features and bridging the modal heterogeneity, while a temporal memory refinement module effectively captures the motion consistency. Although video-based cross-modal person Re-ID is a challenging task, our proposed method achieves remarkable performance, compared with existing state-of-the-art cross-modal approaches.

Acknowledgement

This work was supported in part by Shenzhen Science and Technology Program (RCBS20200714114910193) and the NSFC fund (61906162, 61966021).

References

[1] Abhishek Aich, Meng Zheng, Srikrishna Karanam, Terrence Chen, Amit K Roy-Chowdhury, and Ziyan Wu. Spatio-temporal representation factorization for video-based person re-identification. In *ICCV*, pages 152–162, 2021. 3

[2] Dapeng Chen, Hongsheng Li, Tong Xiao, Shuai Yi, and Xiaogang Wang. Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding. In *CVPR*, pages 1169–1178, 2018. 3

[3] Yanbei Chen, Xiatian Zhu, and Shaogang Gong. Person re-identification by deep learning multi-scale representations. In *ICCV*, pages 2590–2600, 2017. 1

[4] Seokeon Choi, Sumin Lee, Youngeun Kim, Taekyung Kim, and Changick Kim. Hi-cmd: Hierarchical cross-modality disentanglement for visible-infrared person re-identification. In *CVPR*, pages 10257–10266, 2020. 2, 3

[5] Dahjung Chung, Khalid Tahboub, and Edward J Delp. A two stream siamese convolutional neural network for person re-identification. In *ICCV*, pages 1983–1991, 2017. 3

[6] Pingyang Dai, Rongrong Ji, Haibin Wang, Qiong Wu, and Yuyu Huang. Cross-modality person re-identification with generative adversarial training. In *IJCAI*, volume 1, page 2, 2018. 1, 2, 3, 6, 7, 8

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009. 7

[8] Chanho Eom, Geon Lee, Junghyup Lee, and Bumsub Ham. Video-based person re-identification with spatial and temporal memory networks. In *ICCV*, pages 12036–12045, 2021. 3, 5

[9] Zhanxiang Feng, Jianhuang Lai, and Xiaohua Xie. Learning modality-specific representations for visible-infrared person re-identification. *IEEE Transactions on Image Processing*, 29:579–590, 2019. 2, 3

[10] Jiyang Gao and Ram Nevatia. Revisiting temporal modeling for video-based person reid. *arXiv preprint arXiv:1805.02104*, 2018. 3

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *NIPS*, 27, 2014. 3

[12] Xinqian Gu, Hong Chang, Bingpeng Ma, Hongkai Zhang, and Xilin Chen. Appearance-preserving 3d convolution for video-based person re-identification. In *ECCV*, pages 228–243. Springer, 2020. 3

[13] Yi Hao, Nannan Wang, Jie Li, and Xinbo Gao. Hsme: hypersphere manifold embedding for visible thermal person re-identification. In *AAAI*, volume 33, pages 8385–8392, 2019. 2, 3

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 4, 7

[15] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 5

[16] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, pages 7132–7141, 2018. 5

[17] Martin Koestinger, Martin Hirzer, Paul Wohlhart, Peter M Roth, and Horst Bischof. Large scale metric learning from equivalence constraints. In *CVPR*, pages 2288–2295. IEEE, 2012. 1

[18] Diangang Li, Xing Wei, Xiaopeng Hong, and Yihong Gong. Infrared-visible cross-modal person re-identification with an x modality. In *AAAI*, volume 34, pages 4610–4617, 2020. 2

[19] Huafeng Li, Kaixiong Xu, Jinxing Li, Guangming Lu, Yong Xu, Zhengtao Yu, and David Zhang. Dual-stream reciprocal disentanglement learning for domain adaptation person re-identification. *arXiv preprint arXiv:2106.13929*, 2021. 6

[20] Jianing Li, Jingdong Wang, Qi Tian, Wen Gao, and Shiliang Zhang. Global-local temporal representations for video person re-identification. In *ICCV*, pages 3958–3967, 2019. 4

[21] Jianing Li, Shiliang Zhang, and Tiejun Huang. Multi-scale 3d convolution network for video based person re-identification. In *AAAI*, volume 33, pages 8618–8625, 2019. 3

[22] Minxian Li, Xiatian Zhu, and Shaogang Gong. Unsupervised person re-identification by deep learning tracklet association. In *ECCV*, pages 737–753, 2018. 1

[23] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *CVPR*, pages 2285–2294, 2018. 1

[24] Yu-Jhe Li, Yun-Chun Chen, Yen-Yu Lin, Xiaofei Du, and Yu-Chiang Frank Wang. Recover and identify: A generative dual model for cross-resolution person re-identification. In *ICCV*, pages 8090–8099, 2019. 1

[25] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, pages 2197–2206, 2015. 1

[26] Yongguo Ling, Zhun Zhong, Zhiming Luo, Paolo Rota, Shaozi Li, and Nicu Sebe. Class-aware modality mix and center-guided metric learning for visible-thermal person re-identification. In *ACM MM*, pages 889–897, 2020.

[27] Yiheng Liu, Zhenxun Yuan, Wengang Zhou, and Houqiang Li. Spatial and temporal mutual promotion for video-based person re-identification. In *AAAI*, volume 33, pages 8786–8793, 2019. 3

[28] Yan Lu, Yue Wu, Bin Liu, Tianzhu Zhang, Baopu Li, Qi Chu, and Nenghai Yu. Cross-modality person re-identification with shared-specific feature transfer. In *CVPR*, pages 13379–13389, 2020. 2, 3

[29] Niall McLaughlin, Jesus Martinez Del Rincon, and Paul Miller. Recurrent convolutional network for video-based person re-identification. In *CVPR*, pages 1325–1334, 2016. 3

[30] Dat Tien Nguyen, Hyung Gil Hong, Ki Wan Kim, and Kang Ryoung Park. Person recognition system based on a combination of body images from visible light and thermal cameras. *Sensors*, 17(3):605, 2017. 4

[31] Hyunjong Park, Sanghoon Lee, Junghyup Lee, and Bumsub Ham. Learning by aligning: Visible-infrared person re-identification using cross-modal correspondences. In *ICCV*, pages 12046–12055, 2021. 2, 3, 8

[32] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming

Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *NIPS*, 32:8026–8037, 2019. 6

[33] Nan Pu, Wei Chen, Yu Liu, Erwin M Bakker, and Michael S Lew. Dual gaussian-based variational subspace disentanglement for visible-infrared person re-identification. In *ACM MM*, pages 2149–2158, 2020. 2

[34] Lei Qi, Lei Wang, Jing Huo, Luping Zhou, Yinghuan Shi, and Yang Gao. A novel unsupervised camera-aware domain adaptation framework for person re-identification. In *ICCV*, pages 8080–8089, 2019. 6, 7, 8

[35] Xudong Tian, Zhizhong Zhang, Shaohui Lin, Yanyun Qu, Yuan Xie, and Lizhuang Ma. Farewell to mutual information: Variational distillation for cross-modal person re-identification. In *CVPR*, pages 1522–1531, 2021. 8

[36] Guan'an Wang, Tianzhu Zhang, Jian Cheng, Si Liu, Yang Yang, and Zengguang Hou. Rgb-infrared cross-modality person re-identification via joint pixel and feature alignment. In *ICCV*, pages 3623–3632, 2019. 2, 3, 4, 6

[37] Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang. Person re-identification by video ranking. In *ECCV*, pages 688–703. Springer, 2014. 4

[38] Zhixiang Wang, Zheng Wang, Yinqiang Zheng, Yung-Yu Chuang, and Shin'ichi Satoh. Learning to reduce dual-level discrepancy for infrared-visible person re-identification. In *CVPR*, pages 618–626, 2019. 2, 4

[39] Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong, and Jianhuang Lai. Rgb-infrared cross-modality person re-identification. In *ICCV*, pages 5380–5389, 2017. 1, 2, 4

[40] Qiong Wu, Pingyang Dai, Jie Chen, Chia-Wen Lin, Yongjian Wu, Feiyue Huang, Bineng Zhong, and Rongrong Ji. Discover cross-modality nuances for visible-infrared person re-identification. In *CVPR*, pages 4330–4339, 2021. 3, 8

[41] Yu Wu, Yutian Lin, Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang. Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning. In *CVPR*, pages 5177–5186, 2018. 3, 4

[42] Shuangjie Xu, Yu Cheng, Kang Gu, Yang Yang, Shiyu Chang, and Pan Zhou. Jointly attentive spatial-temporal pooling networks for video-based person re-identification. In *ICCV*, pages 4733–4742, 2017. 3

[43] Yang Yang, Jimei Yang, Junjie Yan, Shengcai Liao, Dong Yi, and Stan Z Li. Salient color names for person re-identification. In *ECCV*, pages 536–551. Springer, 2014. 1

[44] Mang Ye, Xiangyuan Lan, and Qingming Leng. Modality-aware collaborative learning for visible thermal person re-identification. In *ACM MM*, pages 347–355, 2019. 2

[45] Mang Ye, Xiangyuan Lan, Qingming Leng, and Jianbing Shen. Cross-modality person re-identification via modality-aware collaborative ensemble learning. *IEEE Transactions on Image Processing*, 29:9387–9399, 2020. 2

[46] Mang Ye, Xiangyuan Lan, Jiawei Li, and Pong Yuen. Hierarchical discriminative learning for visible thermal person re-identification. In *AAAI*, volume 32, 2018. 2

[47] Mang Ye, Xiangyuan Lan, Zheng Wang, and Pong C Yuen. Bi-directional center-constrained top-ranking for visible thermal person re-identification. *IEEE Transactions on Information Forensics and Security*, 15:407–419, 2019. 2, 3

[48] Mang Ye, Weijian Ruan, Bo Du, and Mike Zheng Shou. Channel augmented joint learning for visible-infrared recognition. In *ICCV*, pages 13567–13576, 2021. 3, 8

[49] Mang Ye, Jianbing Shen, David J. Crandall, Ling Shao, and Jiebo Luo. Dynamic dual-attentive aggregation learning for visible-infrared person re-identification. In *ECCV*, pages 229–247. Springer, 2020. 2, 4, 8

[50] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 3, 8

[51] Mang Ye, Zheng Wang, Xiangyuan Lan, and Pong C Yuen. Visible thermal person re-identification via dual-constrained top-ranking. In *IJCAI*, volume 1, page 2, 2018. 2, 3

[52] Wei Zhang, Xiaodong Yu, and Xuanyu He. Learning bidirectional temporal cues for video-based person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(10):2768–2776, 2017. 3

[53] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian. Mars: A video benchmark for large-scale person re-identification. In *ECCV*, pages 868–884. Springer, 2016. 2, 4

[54] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Reidentification by relative distance comparison. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(3):653–668, 2012. 1

[55] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, pages 2223–2232, 2017. 3