IEICE TRANS. INF. & SYST., VOL.E105–D, NO.7 JULY 2022

LETTER

Gray Augmentation Exploration with All-Modality Center-Triplet Loss for Visible-Infrared Person Re-Identification

Xiaozhou CHENG[†,††], **Rui LI**[†], ***Nonmembers***, **Yanjing SUN**[†], ***Member***, **Yu ZHOU**[†a)], ***and*** **Kaiwen DONG**[†], ***Nonmembers***

SUMMARY Visible-Infrared Person Re-identification (VI-ReID) is a challenging pedestrian retrieval task due to the huge modality discrepancy and appearance discrepancy. To address this tough task, this letter proposes a novel gray augmentation exploration (GAE) method to increase the diversity of training data and seek the best ratio of gray augmentation for learning a more focused model. Additionally, we also propose a strong all-modality center-triplet (AMCT) loss to push the features extracted from the same pedestrian more compact but those from different persons more separate. Experiments conducted on the public dataset SYSU-MM01 demonstrate the superiority of the proposed method in the VI-ReID task.

key words: *visible-infrared person re-identification, gray augmentation exploration, AMCT loss, SYSU-MM01*

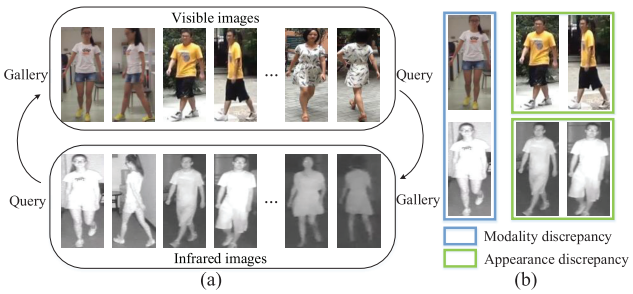

Fig. 1 (a) Illustration of the visible-infrared person re-identification. (b) Examples of modality discrepancy and appearance discrepancy.

1. Introduction

Person re-identification (ReID) aims to retrieve the same person from images captured by disjoint cameras [1]. With the popularization of infrared cameras that are developed to complement the collection defects of visible cameras in dark conditions, the visible-infrared cross-modality ReID (VI-ReID) [2] has been attracting sharply increasing attentions, where the query and gallery images are captured by different modality cameras, as shown in Fig. 1 (a). At present, the advance of VI-ReID lags far behind the VV-ReID in terms of re-identification accuracy [3]. This backwardness is attributed to the modality discrepancy caused by different imaging principles between visible and infrared cameras and the appearance discrepancy originating from visual angles and postures [4], as shown in Fig. 1 (b). All these factors bring huge challenges for VI-ReID.

Up to now, some works have been proposed for the VI-ReID task [2]–[12]. They mainly deal with the above challenges from three aspects, including network designing, metric learning, and image transformation. The network designing based methods [2]–[7] aim to explore superior neural network architectures for better feature learning. The metric learning based methods [8]–[10] promote feature representations by designing the more superior loss function. The image transformation based methods narrow the differences between the query and gallery images through domain conversion or image generation methods for better feature matching [4], [11], [12]. Although these VI-ReID methods have made gradual progress in recent years, there is still a great gap between the performance and satisfactory results.

Modality discrepancy is a key difficulty in visible-infrared person re-identification. The domain adversarial learning [13] is often used to reduce domain discrepancies by encouraging domain-invariant features through the idea of confrontation. It needs to add a domain classifier and a special non-standard gradient reversal layer, which increases the complexity of training. From the perspective of mining modality sharing information like human body structures and reducing modality-specific information such as colors, we propose an effective lightweight gray augmentation exploration (GAE) method to help learn a more focused model. At the same time, to strengthen feature constraints at the feature level, we propose a metric loss called all-modality center-triplet (AMCT) loss, which can effectively reduce the impact of modality discrepancy and appearance discrepancy by feature metric. The contributions of the proposed approach are threefold as follows:

- We put forward an effective GAE method to increase the training data diversity and promote the network to pay more attention to decisive modality sharing information.
- The proposed AMCT loss possesses a strong feature constraint ability. The experiments have demonstrated the superiority compared with the other metric loss forms.
- Extensive experiments show that our approach has

Manuscript received November 5, 2021.
Manuscript revised March 1, 2022.
Manuscript publicized April 6, 2022.
[†]The authors are with School of Information and Control Engineering, China University of Mining and Technology, Xuzhou 221116, China.
[††]The author is with the Sinostell Maanshan General Institute of Mining Research Co., Ltd., Maanshan 243000, China.
a) E-mail: zhouy@cumt.edu.cn (Corresponding author)
DOI: 10.1587/transinf.2021EDL8101

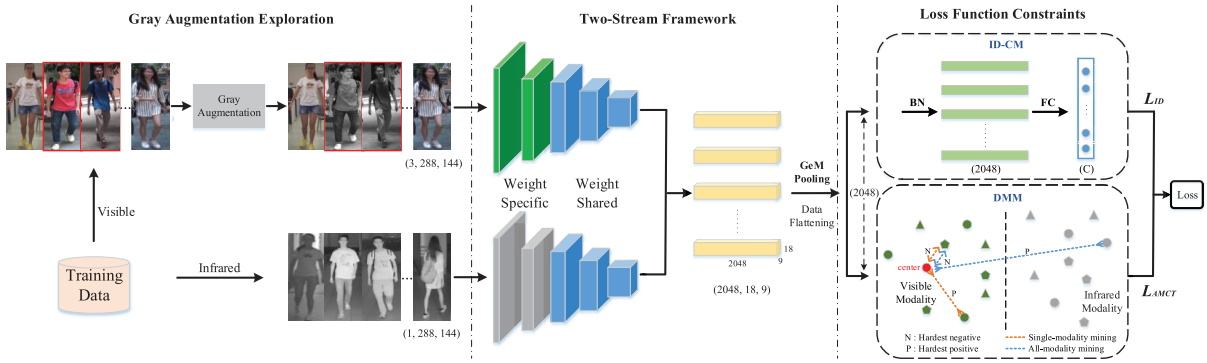

Fig. 2 Overview of our proposed model.

more superior performance and can be a new baseline due to its simplicity and effectiveness.

2. Proposed Method

Our proposed model is shown in Fig. 2. The model consists of three parts, including gray augmentation exploration, two-stream framework, and loss function constraints.

2.1 Gray Augmentation Exploration

VI-ReID can be deemed as a matching or retrieval problem. Naturally, the key of VI-ReID is to measure the similarity of images as accurately as possible. For a pair of visible and infrared images of one pedestrian, they both contain some sharing information like body structures and modality-specific information like colorfulness in visible images, which together affect the human's perception of similarity. An ideal cross-modality ReID model can realize accurate matching by exploring the information shared by different modality images regardless of the modal-specific information. In other words, it is robust to the modality-specific information.

Inspired by the above facts, we propose a GAE method, which urges the network to mine the modality-shared information better and improves its robustness to modality-specific information. Graying is a good choice to remove the color information for more attention to modality sharing information. Graying weights the three color channels R, G, and B of the visible image to generate a gray image, and the weighted values are 0.299, 0.587, and 0.114 respectively. The gray image is a single-channel image, while a visible image contains three channels. We employ the channel expansion strategy to expand the single channel into three channels via a simple replication operation. Same as other visible images, gray images are input into the feature extraction network in the visible domain to extract pedestrian image features. However, we find that graying all images in a mini-batch during training is not the best way, which can be seen from the results in Fig. 3. This phenomenon is predictable, because graying all will result in some vital discriminative information lost. Therefore, an appropriate

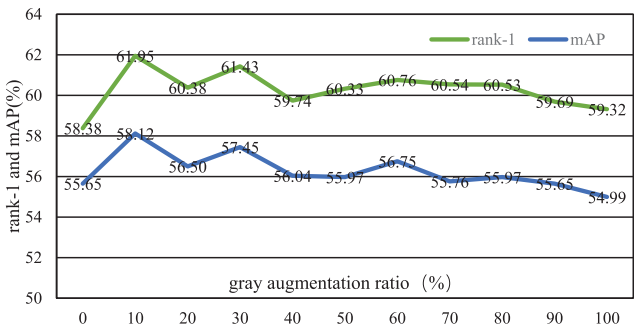

Fig. 3 Rank-1 and mAP values when different augmentation ratios are adopted on the SYSU-MM01 dataset.

ratio of grayscale should be studied. To this end, for the input images in a mini-batch, $t\%$ visible images are randomly selected for graying, and the remaining images are still the original RGB ones. For each iteration, $t\%$ images are randomly chosen again. This method increases the diversity and variability of data, which is helpful to learn a more robust model.

2.2 Two-Stream Framework

This letter uses the two-stream framework as the backbone of the network. The first two stages of Resnet 50 [14] are used for feature extraction, where the parameters are not shared to extract modality respective information. The last three stages share parameters to map features extracted from the previous network to the same feature subspace and explore clues of similarity between modalities. After the last convolutional layer, the generalized-mean pooling (GeM pooling) [3] is appended to transform 3D feature maps into 1D vectors. The vectors are adopted to calculate the AMCT loss and the ID loss after passing through a batch normalization layer and a full connection layer.

2.3 AMCT Loss

There are two modules in the loss function constraints part, i.e., ID classification mudule (ID-CM) and distance metric module (DMM). The ID-CM module is used to distinguish

the features of each identity of pedestrians. The DMM module is used to restrict the distance between features.

In the training, we employ the online batch sampling strategy [3], which can help the model learn images from different modalities of the same pedestrian evenly and effectively avoid the disturbance caused by sample imbalance. P person identities are randomly sampled. For each identity, K visible images and K infrared images are randomly selected as the input of two branches. Through this sampling strategy, the network can learn modality-sharing information and develop the all-modality metric learning.

In the DMM, to improve the constraints of image feature distances, we propose a novel AMCT loss. Different from the common single-modality hard mining sample-triplet metric forms [6], [7] in the VI-ReID task based on sample-anchors to select the hardest parings in only one same modality, the proposed loss aims to learn the centers of the classes of samples and use them instead of individual samples as the anchors to form the hardest triplets, and mines the most difficult positive and negative pairings in all modalities instead of only one modality, as shown in Fig. 2. Specifically, we regard the centers of all identities in each modality of the mini-batch as the anchors, which are calculated as follows:

$$c_V^i = \frac{1}{K}\sum\nolimits_{a=1}^{K} x_{Va}^i, \tag{1}$$

$$c_I^i = \frac{1}{K}\sum\nolimits_{a=1}^{K} x_{Ia}^i, \tag{2}$$

where x_{Va}^i denotes the a^{th} visible image feature of the i^{th} person in the mini-batch, c_V^i denotes the feature center of the i^{th} person in the visible modality, while x_{Ia}^i and c_I^i correspond to the infrared image feature and center. Based on the sampling strategy, the definition of the AMCT loss is as follows.

The first is to calculate the feature distances between the center and the all-modality samples and find the positive center-sample pair with the maximum distance of the same identity and the negative center-sample pair with the minimum distance of different identities. For visible modality:

$$D_{Pos\_V}^i = \max_{\substack{M\in\{V,I\}\\ a=1,\ldots,K}} d\left(c_V^i - x_{Ma}^i\right), \tag{3}$$

$$D_{Neg\_V}^i = \min_{\substack{j\in\{1,\ldots,P|j\neq i\}\\ M\in\{V,I\}\\ a=1,\ldots,K}} d\left(c_V^i - x_{Ma}^j\right), \tag{4}$$

where x_{Ma}^i represents the all-modality image samples with the same identity as c_V^i and x_{Ma}^j represents the all-modality image samples with a different identity from c_V^i. $d(\cdot)$ denotes the Euclidean distance. $D_{Pos\_V}^i$ and $D_{Neg\_V}^i$ are the most difficult positive and negative center-sample pairs of c_V^i , respectively. Similarly, for the infrared modality center c_I^i :

$$D_{Pos\_I}^i = \max_{\substack{M\in\{V,I\}\\ a=1,\ldots,K}} d\left(c_I^i - x_{Ma}^i\right), \tag{5}$$

$$D_{Neg\_I}^i = \min_{\substack{j\in\{1,\ldots,P|j\neq i\}\\ M\in\{V,I\}\\ a=1,\ldots,K}} d\left(c_I^i - x_{Ma}^j\right). \tag{6}$$

Then, the most difficult positive and negative pairs form a triple form,

$$L_{AMCT\_V} = \frac{1}{P}\sum\nolimits_{i=1}^{P}[D_{Pos\_V}^i - D_{Neg\_V}^i + m]_+, \tag{7}$$

$$L_{AMCT\_I} = \frac{1}{P}\sum\nolimits_{i=1}^{P}[D_{Pos\_I}^i - D_{Neg\_I}^i + m]_+, \tag{8}$$

where m is a margin parameter. $[X]_+$ means to take the larger value between X and 0. Finally,

$$L_{AMCT} = L_{AMCT\_V} + L_{AMCT\_I}. \tag{9}$$

Compared to the existing metric forms, the proposed AMCT loss has three key contributions. The all-modality mining can not only dig correlations between two modalities but also improve the training efficiency benefiting from more representative training data. Moreover, the center-anchor triplets can make the compactness of the same identity samples and the separation of different identity samples have a clearer goal. Compared to traditional sample-anchor triplets, the computational cost in our method reduces from $2PK$ to $2P$.

ID loss used in the ID-CM and the proposed AMCT loss are employed to supervise the network training jointly to learn a model with strong discrimination ability. For a given image, its ID prediction logic of all identities generated by the network and softmax operation is p_i, while the distribution of its real ID label is q_i. The calculation of ID loss is in the following:

$$L_{ID} = L_{ID\_V} + L_{ID\_I} = \sum\nolimits_{i=1}^{2PK} -q_i \log(p_i) \tag{10}$$

The total loss of the whole network:

$$L_{ALL} = L_{ID} + L_{AMCT} \tag{11}$$

3. Experiments

3.1 Experimental Settings

3.1.1 Datasets and Metrics

Our experiments are based on the standard benchmark for the VI-ReID task, named SYSU-MM01 [2]. It provides 491 valid pedestrian identities captured from six cameras (four visible ones and two infrared ones), including 287,628 visible images and 15,792 infrared images. Person images are captured in both indoor and outdoor environments.

On the SYSU-MM01 dataset, there are two search modes, i.e., all-search mode and indoor-search mode. Following the most popular protocol [2], [9], [12], the cumulative matching characteristics (CMC) and mean average precision (mAP) are adopted as the evaluation metrics, where higher values indicate the better retrieval performance.

Table 1 Comparison to the state-of-the-arts on the SYSU-MM01 dataset. Re-identification rates at rank-r (%) and mAP (%)

All-search					
Method	Venue	rank-1	rank-10	rank-20	mAP
Zero-Pad [2]	ICCV17	14.80	54.12	71.33	15.95
HCML [5]	AAAI18	14.32	53.16	69.17	16.16
D2RL [4]	CVPR19	28.90	70.60	82.40	29.20
MAC [6]	MM19	33.26	79.04	90.09	36.22
EDFL [7]	Neuro20	36.94	85.42	93.22	40.77
XIV [12]	AAAI20	49.92	89.79	95.96	50.73
HC [8]	Neuro20	56.96	91.50	96.82	54.95
DDAG [9]	ECCV20	54.75	90.39	95.81	53.02
AGW [3]	TPAMI21	47.50	84.39	92.14	47.65
Proposed	-	**61.95**	**93.37**	**97.65**	**58.12**
Indoor-search					
Zero-Pad [2]	ICCV17	20.58	68.38	85.79	26.92
HCML [5]	AAAI18	24.52	73.25	86.73	30.08
MAC [6]	MM19	36.43	62.36	71.63	37.03
HC [8]	Neuro20	59.74	92.07	96.22	64.91
DDAG [9]	ECCV20	61.02	94.06	98.41	67.98
AGW [3]	TPAMI21	54.17	91.14	95.98	62.97
Proposed	-	**63.68**	**94.98**	**98.11**	**69.91**

Table 2 Comparisons of different losses performance on SYSU-MM01. Rank-1 accuracy (%) and mAP (%) are reported.

Method	All-search		Indoor-search	
	rank-1	mAP	rank-1	mAP
Only ID	31.89	27.27	37.26	43.47
ID+SMST	45.96	46.73	48.58	57.99
ID+SMCT	49.87	48.34	49.99	59.07
ID+AMST	57.27	55.39	59.68	67.54
ID+AMCT (Proposed)	**58.38**	**55.65**	**61.96**	**69.01**

Table 3 Performance of the proposed method with different number of shared stages on SYSU-MM01. Rank-1 accuracy (%) and mAP (%) are reported.

Shared stage	All-search		Indoor-search	
	rank-1	mAP	rank-1	mAP
Stage0-Stage4	60.40	56.67	**63.80**	69.72
Stage1-Stage4	59.02	55.80	59.89	67.11
Stage2-Stage4	**61.95**	**58.12**	63.68	**69.91**
Stage3-Stage4	59.69	56.42	60.91	68.37
Stage4	54.40	52.95	57.12	65.29
No Stage	46.41	46.56	50.15	59.73

3.1.2 Implementation Details

Our method is implemented with the Pytorch framework. Like most existing VI-ReID works, we adopt Resnet50 as the backbone network for fair comparison. The pre-trained ImageNet parameters are adopted for the network initialization. In the training phase, all input images are resized to 288×144 and padded with 10. Then, they are randomly cut into 288×144 and flipped horizontally for data augmentation. We use the stochastic gradient descent (SGD) optimizer for optimization. The momentum parameter is set to 0.9. For the online sampling strategy, $P = 4$, $K = 8$ are set. In addition, the margin m is set to 0.3. 60 epochs are trained. All the training and testing procedures are completed on a server with Tesla v100a-sxm2.

3.2 Comparison to the State-of-the-Arts

In this part, we present the performance of our proposed method with the state-of-the-arts. Table 1 shows the results on the SYSU-MM01 dataset, where the best results are marked in boldface. Our method obtains the highest performance in terms of the most significant indicators, i.e. rank-1 and mAP. In conclusion, our method is highly competitive to existing methods and reaches the current leading level.

3.3 Ablation Experiments

3.3.1 The Effect of AMCT Loss

To testify the effect of the proposed AMCT loss, we test the performance of the proposed method by removing it or replacing it with the SMST (single-modality sample-triplet) [7], SMCT (single-modality center-triplet) and AMST (all-modality sample-triplet) loss, respectively in the DMM. From Table 2, it can be observed that the performance drops sharply if the AMCT loss is not employed. Besides, the network supervised by the AMCT loss achieves the best results whether in the all-search mode or the indoor-search mode. To sum up, the AMCT loss proposed by us achieves the best results and can effectively enhance the constraints on feature distances in the VI-ReID task. This benefits from its stronger abilities of feature aggregation within class and feature separation among different classes.

3.3.2 The Effect of GAE

We further testify the optimal gray augmentation ratio by changing it from 0% to 100% with the step of 10%. The results on the more challenging all-search mode of the SYSU-MM01 dataset are shown in Fig. 3. As can be seen, when the ratio of gray augmentation is 10%, the best performance is obtained. Moreover, compared with no gray augmentation, our method achieves an improvement of 3.57% and 2.47% in terms of rank-1 and mAP respectively. The experimental results indicate the effectiveness of our GAE method. Moderate gray transformation is helpful to reduce the discrepancy between modalities and the influence of modality-specific information.

3.3.3 The Ablation of the Network Structure

In this subpart, we further conduct experiments to investigate the effect of different number of shared stages on the performance of the proposed method. The results can be seen in Table 3, from which we can see that when different stages are shared, the performance is different. When Stage2-Stage4 are shared, our method achieves the highest recognition accuracy on most indicators. When all stages

are not shared, the proposed method has the worst performance, which indicates that two modality images share some clues to a certain extent.

4. Conclusion

In this letter, we have presented a novel VI-ReID method based on gray augmentation exploration and the all-modality center-triplet loss. GAE reduces the discrepancies between the two modalities at the image level. At the feature level, the constraints within and among classes are strengthened by our AMCT loss, and better performances of VI-ReID are achieved. Experimental results on the public database demonstrate the advantages of the proposed metric over the relevant state-of-the-arts.

Acknowledgments

This work is supported by the Natural National Science Foundation of China (61902404, 62001475, 51734009, 61771417, 61873246).

References

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

[2] A. Wu, W.-S. Zheng, H.-X. Yu, S. Gong, and J. Lai, "Rgb-infrared cross-modality person re-identification," Proc. IEEE Int. Conf. Comput. Vis., pp.5380–5389, 2017.

[3] M. Ye, J.-B. Shen, G.-J. Lin, T. Xiang, L. Shao, and S.C.H. Hoi, "Deep learning for person re-identification: A survey and outlook," IEEE Trans. Pattern Anal. Mach. Intell., vol.44, no.6, pp.2872–2893, 2021. DOI: 10.1109/TPAMI.2021.3054775.

[4] Z. Wang, Z. Wang, Y. Zheng, Y.-Y. Chuang, and S. Satoh, "Learning to reduce dual-level discrepancy for infrared-visible person reidentification," Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp.618–626, 2019.

[5] M. Ye, X. Lan, J. Li, and P.C. Yuen, "Hierarchical discriminative learning for visible thermal person re-identification," Proc. AAAI, pp.7501–7508, 2018.

[6] M. Ye, X.-Y. Lan, and Q.-M. Leng, "Modality-aware collaborative learning for visible thermal person re-identification," ACM Multimedia, pp.347–355, 2019.

[7] H.-J. Liu, J. Cheng, W. Wang, Y. Su, and H. Bai, "Enhancing the discriminative feature learning for visible-thermal cross-modality person re-identification," Neurocomputing, vol.398, pp.11–19, 2020.

[8] Y. Zhu, Z. Yang, L. Wang, S. Zhao, X. Hu, and D. Tao, "Hetero-center loss for cross-modality person re-identification," Neurocomputing, vol.386, pp.97–109, 2020.

[9] M. Ye, J. Shen, D.J. Crandall, L. Shao, and J.-B. Luo, "Dynamic Dual-Attentive Aggregation Learning for Visible-Infrared Person Re-Identification," Proc. IEEE Int. Conf. Eur. Conf. Comput. Vis., pp.229–247, 2020.

[10] H.-R. Ye, H. Liu, F.-Y. Meng, and X. Li, "Bi-directional Exponential Angular Triplet Loss for RGB-Infrared Person Re-Identification," IEEE Trans. Image Process., vol.30, pp.1583–1595, 2020. DOI: 10.1109/TIP.2020.3045261.

[11] G. Wang, T. Zhang, J. Cheng, S. Liu, Y. Yang, and Z. Hou, "Rgb-infrared cross-modality person re-identification via joint pixel and feature alignment," Proc. IEEE Int. Conf. Comput. Vis., pp.3623–3632, 2019.

[12] D.-G. Li, X. Wei, X. Hong, and Y. Gong, "Infrared-visible cross-modal person re-identification with an X modality," Proc. AAAI, pp.4610–4617, 2020.

[13] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," Journal of Machine Learning Research, vol.17, no.1, pp.2096–2030, 2016.

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