# Diffusion Augmentation and Pose Generation Based Pre-Training Method for Robust Visible-Infrared Person Re-Identification

Rui Sun, *Member, IEEE*, Guoxi Huang, Ruirui Xie, Xuebin Wang, and Long Chen

***Abstract*—Cross-Modal Visible-Infrared Person Re-identification (VI-REID) constitutes a vital application for constructing all-time surveillance systems. However, the current VI-REID model exhibits significant performance deterioration in noisy environments. Existing algorithms endeavor to mitigate this challenge through fine-tuning stages. We contend that, in contrast to fine-tuning stages, the pre-training phase can effectively exploit the attributes of extensive unlabeled data, thereby facilitating the development of a robust VI-REID model. Therefore, in this paper, we propose a pre-training method for VI-REID based on Diffusion Augmentation and Pose Generation (DAPG), aiming to enhance the robustness and recognition rate of VI-REID models in the presence of damaged scenes. Multiple transfer experiments on the SYSU-MM01 and RegDB datasets demonstrate that our method outperforms existing self-supervised methods, as evidenced by the results.**

***Index Terms*—Person re-identification, visible-infrared, pre-training, self-supervised, corruption robustness.**

## I. Introduction

VISIBLE-INFRARED Person Re-identification (VI-ReID) [1] is an image retrieval task across scenes and cameras, aiming to determine whether a target pedestrian appearing in one camera can be found in multiple non-overlapping visual regions. It has been widely applied in constructing all-time surveillance systems. However, in practical scenarios, the performance of VI-ReID models suffers severe degradation due to complex noise present in real-world environments.

Currently, in the realm of person re-identification, there exist several studies addressing the robustness to image corruption [2], [3], exemplified by works like CIL [4] and PMWGCN [5]. These studies typically commence at the fine- tuning stage, primarily employing data augmentation [3] to simulate the data distribution present in real-world scenarios, thereby enhancing the robustness of model in corrupted settings. However, due to the limited quantity of data during the fine-tuning stage, solely relying on data augmentation on the fine-tuning dataset imposes significant limitations. In fact, methodologies aimed at enhancing robustness through the fine-tuning stage may result in a decline in the overall recognition performance of the model. As shown in Fig. 1, CIL and PMWGCN methods improve the recognition rate of the model in the corrupted scenario, but decrease the recognition rate in the clean dataset.

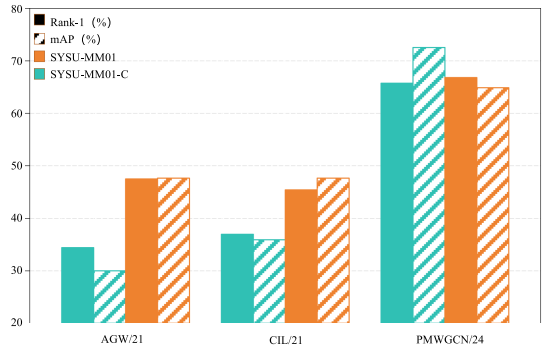


Fig. 1. Recognition rate diagram of AGW [1], CIL [2], PMWGCN [3].

This paper embarks from the perspective of pre-training, aiming to address the decline of accuracy in damaged noisy environments, and to enhance the robustness of model by leveraging the advantages of large-scale data. We proposed a pre-training method based on Diffusion Augmented and Pose Generation (DAPG) to facilitate the learning of damage-invariant feature representations. Finally, the effectiveness of the proposed algorithm is validated through experiments conducted on datasets such as SYSU-MM01 [6] and RegDB [7].

## II. Method

Our contrastive pre-training algorithm based on DAPG is illustrated in Fig. 2. It is divided into two main stages: robustness training and aligning the training objectives of pre-training and fine-tuning. In the first stage, utilizing our constructed contrastive diffusion framework, the model is encouraged to learn noise-resistant capabilities to enhance its robustness. In the second stage, we align contrastive learning tasks [8], [9], [10], [11], [12] with person re-identification tasks through pose generation, mitigating the loss of transferable knowledge resulting from significant disparities between pre-training and fine-tuning tasks.

Received 28 May 2024; revised 13 September 2024; accepted 18 September 2024. Date of publication 23 September 2024; date of current version 3 October 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62302142 and Grant 61876057, and in part by the Natural Science Foundation of Anhui Province under Grant 2208085MF158. The associate editor coordinating the review of this article and approving it for publication was Dr. Anurag Kumar. *(Corresponding author: Rui Sun.)*

Rui Sun is with the Key Laboratory of Knowledge Engineering With Big Data (Ministry of Education), School of Computer and Information, Anhui Province Key Laboratory of Industry Safety and Emergency Technology, Hefei University of Technology, Hefei 230009, China (e-mail: sunrui@hfut.edu.cn).

Guoxi Huang, Ruirui Xie, Xuebin Wang, and Long Chen are with the School of Computer and Information, Anhui Province Key Laboratory of Industry Safety and Emergency Technology, Hefei University of Technology, Hefei 230009, China.

Digital Object Identifier 10.1109/LSP.2024.3466792

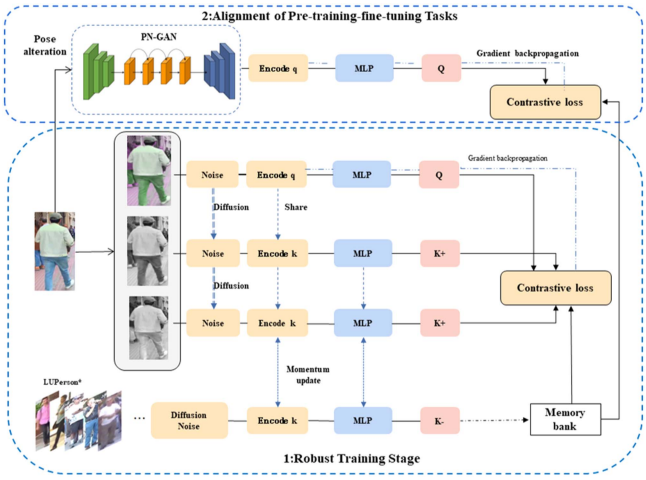


Fig. 2. Schematic diagram of DAPG.

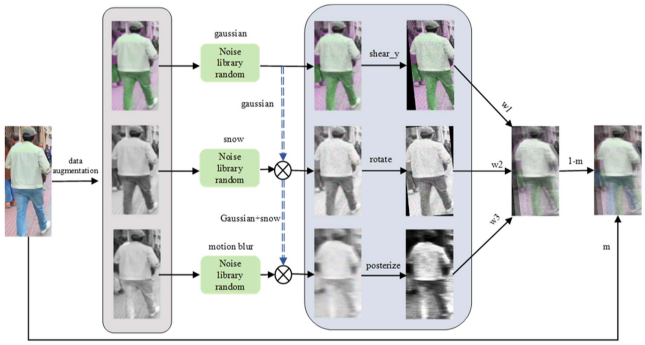


Fig. 3. Flowchart of diffusion noise.

### A. Robustness Training Based on Diffusion Augmentation

Following the noise definitions in ImageNet-C [13], we compile a noise library consisting of 15 types of algorithmically generated image corruptions, including motion blur, Gaussian noise, foggy conditions, snowy conditions, and others.

First, we randomly select an image from the dataset and generate three positive sample images by data augmentation. Among them, we randomly select one image as the anchor image, then randomly select one noise from the noise library, apply it to the anchor image, and obtain the so-called anchor noise image $x^i_{q,noise}$. Then, we apply the second noise on top of the first noise to generate the second noise image $x^i_{1,noise}$. Similarly, the third positive sample image adds the third type of noise on top of the first two fixed noises, resulting in the third noise image $x^i_{2,noise}$. Based on these three noise images, combined with the augMix [14] algorithm, we construct the Mix image. Thus, the creation of positive sample noise groups and negative sample noise diffusion follows the same logic. Through this process, this chapter successfully implements the creation of positive sample noise groups and negative sample noise groups, with these images being subjected to different levels of noise. The specific noise diffusion process is illustrated in Fig. 3.

### B. Task Alignment Based on Pose Generation

There are significant differences in the training perspectives between contrastive learning-based pre-training methods and visible-infrared person re-identification tasks. Specifically, contrastive learning treats multiple augmented views of a single image as a unique positive sample, thus driving the model to learn a unique representation for each image. However, the goal in visible-infrared person re-identification is to distinguish the identity of people rather than learning a unique representation for every instance. Therefore, person re-identification tasks expect images from the same person had similar representations. Additionally, pose changes [15] introduce noise in person re-identification task, which is from structural differences, differences in camera angles and in cross-modal images.

To address the above issues, we propose a pre-training fine-tuning alignment method based on pose generation. The pre-training fine-tuning alignment phase occurs after robustness training and before transfer fine-tuning. It aims to alleviate knowledge loss during the transfer from contrastive learning-based pre-training tasks to visible-infrared pedestrian re-identification tasks, as well as to mitigate structural noise issues inherent in pedestrian recognition.

In this phase, we introduce the PN-GAN [16] module into the data augmentation process of contrastive learning to perform changes in pedestrian poses, thereby correcting the original objective of learning a unique representation for each image in contrastive learning, and instead encouraging the model to learn identity attributes of pedestrian images. Additionally, we also perform channel enhancement on the modified poses to address the modality differences between the pre-training and fine-tuning stages. The input to the PN-GAN model consists of an original pedestrian image $x_{ori}$ and a desired target pose image $x_{target-pose}$, while the output is the pedestrian image $x_{pose-changed}$, under the target pose. To ensure the quality of generation, we choose the Market-1501 [17] dataset with higher-quality pedestrian images as the pre-training data. During training, we do not use labels, and we specify four uniform pose images as the generation targets, as illustrated in Fig. 4.

During the actual training process, we randomly select one pose from the four target poses as the generation target. Then, we treat $x_{ori}$ as the anchor image, $x_{target-pose}$ as the positive sample image, and randomly select other samples from the Market-1501 dataset as negative samples to complete self-supervised training.

### C. Loss Function

The comprehensive loss function is formulated as follows:

$$L_{total} = \partial L_{MultiContrastive}(X^+, X^-) + (1-\partial)\text{Constraint}(x^1_{noise}, x^2_{noise}, x^3_{mix}) \quad (1)$$

Where $L_{MultiContrastive}$ represents the contrastive cross-entropy loss for multiple positive samples, Constraint stands for the consistency constraint loss, and $\partial$ denotes the weight coefficients between losses.

In order to accommodate the format of multiple positive samples, we have modified the traditional contrastive cross-entropy loss. The modified loss function for multiple positive samples is represented as follows:

$$L_{MultiContrastive} = -\log \frac{\sum_{i=1}^{3} exp(q \cdot {k_i}^+/\tau)}{\sum_{i=1}^{3} exp(q \cdot {k_i}^+/\tau) + \sum_{j=1}^{m-1} exp(q \cdot k_j^-/\tau)} \quad (2)$$

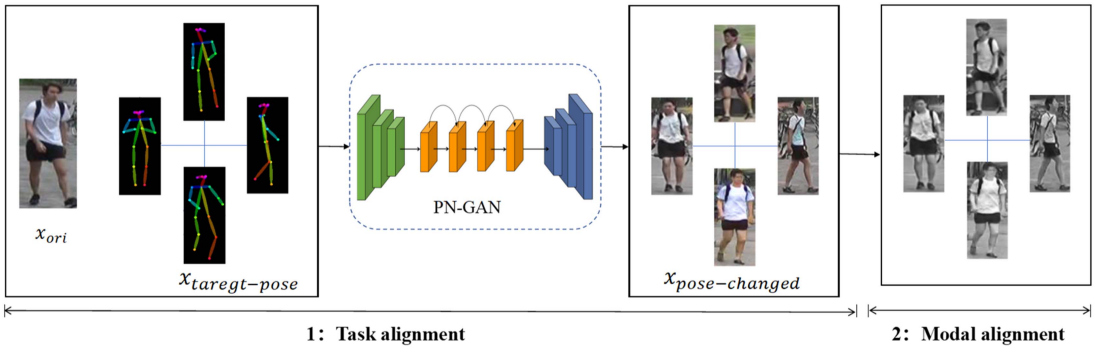


Fig. 4. Task alignment diagram.

$k_i{}^+$ represents the features extracted by the encoder from the augmented images of positive sample group $X^+$, $k_j^-$ represents the features extracted by the encoder from the augmented images of positive sample group $X^-$, $\tau$ denotes the temperature hyperparameter, which controls the focus of the contrastive loss on hard samples. Generally, a higher value of $\tau$ indicates a higher focus of the model on hard negative samples, while excessively large $\tau$ may reduce the discriminative ability between positive and negative samples.

We use Jensen-Shannon divergence (JS divergence) to construct consistency constraint loss, primarily aimed at encouraging a tighter distribution of similarities among positive samples. This prevents excessive differences between positive samples, thus enhancing the stability of the model during training. The consistency constraint loss is specifically represented as follows:

$$\begin{aligned}&\text{Constraint}\left(x^1_{noise}, x^2_{noise}, x^3_{mix}\right)\\&= JS\left(p_{x^1_{noise}}, p_{x^2_{noise}}, p_{x^3_{mix}}\right)\end{aligned} \tag{3}$$

Where $p_{x^1_{noise}}$ is the probability distribution of augmented images, calculated using the softmax [18] function. JS divergence is an improved measure of KL divergence, which can quantify the similarity between multiple probability distributions. The calculation process of JS divergence is represented as the following equation:

$$M = (p_{x^1_{noise}}, p_{x^2_{noise}}, p_{x^3_{mix}})/3 \tag{4}$$

$$\begin{aligned}&\text{JS}\left(p_{x^1_{noise}}, p_{x^2_{noise}}, p_{x^3_{mix}}\right) = \frac{1}{3}\left(\text{KL}(p_{x^1_{noise}}|M)\right.\\&\left.+\text{KL}(p_{x^2_{noise}}|M) + \text{KL}(p_{x^3_{mix}}|M)\right)\end{aligned} \tag{5}$$

## III. Experiments

### A. Datasets and Evaluation Metrics

During the robustness training phase and the pre-training fine-tuning alignment phase of the training process, we utilize the LUPerson [19] and Market-1501 [17] datasets for pre-training, without leveraging the labels from Market-1501. In the testing phase, we evaluate our model on four datasets: RegDB [7], SYSU-MM01 [6], RegDB-C, and SYSU-MM01-C. RegDB-C and SYSU-MM01-C are corrupt versions of the original RegDB and SYSU-MM01 datasets, respectively, with added noise while maintaining a similar data structure.

LUPerson consists of over 70000 videos obtained from YouTube by searching for "city name + street view" format. These videos are then processed using YOLO-V5 [20] to detect and crop pedestrian images. The dataset comprises over 20000 unlabelled person images representing more than 200000 pedestrian identities. Market-1501 is a dataset collected by Tsinghua University, containing 32668 visible light pedestrian images belonging to 1501 pedestrian identities.

RegDB is a small-scale dataset captured by a dual-camera system, consisting of 8024 images from 412 identities. On average, each identity has 10 visible light images and 10 infrared images. SYSU-MM01, a prominent dataset for cross-modality pedestrian re-identification, is captured by four visible light cameras and two near-infrared cameras. It comprises 491 pedestrian identities with 287628 visible light RGB images and 15792 near-infrared images.

### B. Experimental Setting

The proposed DAPG method is implemented by Pytorch. The training process is generally divided into two stages. The robustness training stage involves training for 300 epochs, while the pretraining-finetuning alignment stage involves training for 200 epochs. The training process utilizes the Adam [21] optimizer for optimization, with an initial learning rate set to 0.0075. According to experience and many experiments, the hyperparameters $\partial$ and $\tau$ are set to 0.5 and 0.7, respectively.

During the training stage, all images are resized to $256 \times 128$. Additionally, random cropping, horizontal flipping, and grayscale conversion are employed as auxiliary augmentation methods.

### C. Comparative Experimental Results

In this section, we conduct tests by transferring pre-trained models to two representative mainstream visible-infrared person re-identification downstream tasks, AGW [1] and CIL [4]. The tests are on clean datasets RegDB and SYSU-MM01, and on noisy datasets RegDB-C and SYSU-MM01-C. We compare the results with those of common pre-training methods to discuss the superiority of our algorithm based on DAPG. The compared pre-training algorithms fall into two categories: supervised and unsupervised. The supervised algorithms involve models from [22] and PNL [23], while the unsupervised algorithms include BYOL [13], SimCLR [11], SSL [24] and VTBR [25]. Additionally, model in [22], We denote it as IN Sup, is trained on the ImageNet dataset, we modified it to satisfy the ReID task

TABLE I
RESULTS OF PRE-TRAINING METHODS ON SYSU-MM01 AND REGDB

| Method | Pre-training Method | All-search | | Indoor-search | | Visible-to-Thermal | | Thermal-to-Visible | |
|---|---|---|---|---|---|---|---|---|---|
| | | Rank 1 | mAP | Rank 1 | mAP | Rank 1 | mAP | Rank 1 | mAP |
| CIL | IN Sup | 45.41 | 47.64 | 50.98 | 60.45 | 74.96 | 69.75 | 74.95 | 69.21 |
| | PNL | 54.94 | 56.02 | 57.85 | 66.92 | 77.02 | 71.83 | 77.91 | 72.47 |
| | BYOL | 48.23 | 49.24 | 52.47 | 63.58 | 75.58 | 70.74 | 75.74 | 70.83 |
| | SSL | 46.18 | 46.89 | 52.07 | 61.53 | 72.69 | 68.48 | 73.28 | 69.81 |
| | VTBR | 51.68 | 53.42 | 53.06 | 64.81 | 77.02 | 71.18 | 78.52 | 70.43 |
| | SimCLR | 46.78 | 47.88 | 51.23 | 61.85 | 73.14 | 68.46 | 73.89 | 67.97 |
| | Ours | 56.67 | 59.36 | 59.27 | 70.76 | 77.85 | 71.23 | 78.95 | 71.52 |
| AGW | IN Sup | 47.50 | 47.65 | 54.17 | 63.97 | 70.05 | 66.37 | 70.49 | 65.90 |
| | PNL | 55.36 | 55.45 | 57.85 | 66.92 | 69.75 | 66.12 | 69.14 | 66.59 |
| | BYOL | 51.14 | 50.14 | 55.68 | 65.14 | 70.17 | 66.61 | 70.37 | 63.22 |
| | SSL | 47.21 | 48.02 | 52.86 | 63.76 | 70.93 | 68.29 | 71.56 | 68.86 |
| | VTBR | 53.26 | 52.93 | 56.52 | 66.31 | 73.54 | 68.75 | 74.87 | 69.63 |
| | SimCLR | 48.27 | 48.28 | 54.69 | 63.84 | 67.49 | 63.35 | 68.18 | 61.67 |
| | Ours | 56.14 | 55.03 | 58.33 | 69.52 | 74.26 | 70.44 | 76.23 | 65.85 |

TABLE II
RESULTS OF PRE-TRAINING METHODS ON SYSU-MM01-C AND REGDB-C

| Method | Pre-training Method | All-search | | Indoor-search | | Visible-to-Thermal | | Thermal-to-Visible | |
|---|---|---|---|---|---|---|---|---|---|
| | | Rank 1 | mAP | Rank 1 | mAP | Rank 1 | mAP | Rank 1 | mAP |
| CIL | IN Sup | 36.95 | 35.92 | 40.73 | 48.65 | 52.55 | 49.76 | 67.17 | 47.90 |
| | PNL | 43.59 | 52.75 | 45.70 | 54.08 | 62.73 | 59.96 | 70.38 | 68.49 |
| | BYOL | 38.01 | 37.36 | 42.27 | 50.22 | 54.12 | 50.53 | 68.24 | 53.25 |
| | SSL | 36.92 | 36.22 | 40.07 | 49.16 | 51.83 | 50.07 | 53.24 | 52.46 |
| | VTBR | 42.87 | 43.36 | 47.02 | 53.72 | 62.37 | 59.46 | 69.62 | 62.09 |
| | SimCLR | 37.39 | 36.22 | 41.21 | 49.08 | 52.34 | 49.02 | 66.43 | 50.71 |
| | Ours | 48.36 | 55.88 | 49.27 | 57.21 | 67.41 | 60.04 | 72.17 | 56.63 |
| AGW | IN Sup | 34.42 | 29.99 | 35.24 | 41.57 | 45.44 | 43.09 | 67.54 | 41.37 |
| | PNL | 44.50 | 49.47 | 47.68 | 54.92 | 55.28 | 52.41 | 64.56 | 61.74 |
| | BYOL | 36.28 | 32.14 | 33.92 | 41.07 | 47.14 | 44.52 | 68.75 | 44.13 |
| | SSL | 38.07 | 38.45 | 39.37 | 49.76 | 60.07 | 58.34 | 62.21 | 59.82 |
| | VTBR | 43.18 | 46.91 | 44.05 | 53.29 | 63.84 | 60.29 | 71.47 | 63.29 |
| | SimCLR | 35.74 | 30.25 | 42.58 | 50.27 | 45.27 | 42.88 | 66.21 | 41.07 |
| | Ours | 47.63 | 51.93 | 45.95 | 55.40 | 65.53 | 55.73 | 73.16 | 55.36 |

and tested it on the SYSU-MM01 and RegDB datasets, with the results presented in the Tables I and II.

We employ six pre-trained models for each fine-tuning method to facilitate comparison. Experimental results indicate that in the All-search mode of SYSU-MM01, our proposed method achieves a Rank 1 accuracy 11.26% higher when transferred to the CIL model compared to supervised pre-training on ImageNet. Similarly, when transferred to the AGW model, the Rank 1 accuracy is 8.64% higher. Moreover, in the Visible-to-Thermal mode of the RegDB dataset, our method achieves a 2.89% higher Rank 1 accuracy when transferred to the CIL model compared to supervised pre-training on ImageNet. This suggests that our pre-training approach enhances the generalization ability of model, making it perform better when handling new, unknown, or challenging data.

Table II present the experimental results of our pre-training method compared to other pre-training methods on the corrupted datasets SYSU-MM01-C and RegDB-C. The experimental results show that in the All-search mode of the SYSU-MM01-C dataset, our pre-training method achieves a Rank 1 recognition rate 11.41% higher when transferred to the CIL model compared to its original model pre-trained on ImageNet. The conclusion can be drawn that even on damaged datasets, our algorithm maintains superior performance compared to other pre-training methods, demonstrating its advantage in robustness.

### D. Ablation Study

Table III presents the effectiveness of each part of our method on the corrupted datasets SYSU-MM01-C. As shown in index 1 and 3 of Table III, our Robust Training Stage provides a solid baseline for ReID, our contrastive loss can optimize the model's learning performance and enhance its robustness. Moreover, index 2 and 4 indicate that conducting only the fine-tuning stage cannot enable the model to achieve satisfactory generalization capabilities. Integrating the first-stage Robust Training can further improve generalization abilities and robustness of the model.

TABLE III
ABLATION STUDY RESULTS BASED ON AGW AND TESTED ON SYSU-MM01-C

| Index | Stage1 | Stage2 | $Loss_{total}$ | All Search | | Indoor Search | |
|---|---|---|---|---|---|---|---|
| | | | | Rank-1 | mAP | Rank-1 | mAP |
| 1 | ✓ | | | 33.49 | 30.62 | 34.81 | 34.33 |
| 2 | | ✓ | | 28.05 | 26.53 | 29.67 | 29.16 |
| 3 | ✓ | | ✓ | 39.06 | 42.87 | 38.81 | 41.94 |
| 4 | ✓ | ✓ | ✓ | 47.63 | 51.93 | 45.95 | 55.40 |

## IV. CONCLUSION

In this paper, we propose a robust visible-infrared cross-modal person re-identification pre-training method based on Diffusion Augmentation and Pose Generation (DAPG). The method is mainly divided into two stages: the first stage is the robust training stage, in which we propose a diffusion noise augmentation strategy to enrich the diversity of training noise samples as much as possible. The second stage is the pre-training-fine-tuning task alignment stage, which aims to reduce the difference between pre-training methods and downstream task methods. To achieve this, we propose a posture-based alignment strategy in this stage, correcting the objectives of traditional contrastive learning. Finally, through comparative experiments and ablation experiments on multiple datasets, the effectiveness of the proposed pre-training method is demonstrated.

## REFERENCES


[1] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. H. Hoi, "Deep learning for person re-identification: A survey and outlook," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 6, pp. 2872–2893, Jun. 2022.

[2] F. Yang et al., "Towards robust person re-identification by defending against universal attackers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 4, pp. 5218–5235, Apr. 2023.

[3] A. Josi, M. Alehdaghi, R. M. Cruz, and E. Granger, "Multimodal data augmentation for visual-infrared person ReID with corrupted data," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2023, pp. 32–41.

[4] M. Chen, Z. Wang, and F. Zheng, "Benchmarks for corruption invariant person re-identification," *Neural Information Processing Systems Track on Datasets and Benchmarks,* 2021.

[5] R. Sun, L. Chen, L. Zhang, R. Xie, and J. Gao, "Robust visible-infrared person re-identification based on polymorphic mask and wavelet graph convolutional network," *IEEE Trans. Inf. Forensics Secur.*, vol. 19, pp. 2800–2813, 2024.

[6] A. Wu, W.-S. Zheng, H.-X. Yu, S. Gong, and J. Lai, "RGB-infrared cross-modality person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017: pp. 5390–5399.

[7] D. T. Nguyen, H. G. Hong, K. W. Kim, and K. R. Park, "Person recognition system based on a combination of body images from visible light and thermal cameras," *Sensors*, vol. 17, no. 3, pp. 1–29, Mar. 2017.

[8] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020: pp. 9729–9738.

[9] S. Gwon, S. Kim, and K. Seo, "Balanced and essential modality-specific and modality-shared representations for visible-infrared person re-identification," *IEEE Signal Process. Lett.*, vol. 31, pp. 491–495, 2024.

[10] W. Hou, W. Wang, Y. Yan, D. Wu, and Q. Xia, "A three-stage framework for video-based visible-infrared person re-identification," *IEEE Signal Process. Lett.*, vol. 31, pp. 1254–1258, 2024.

[11] J. B. Grill et al., "Bootstrap your own latent-a new approach to self-supervised learning," *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 21271–21284, 2020.

[12] J. Jiang and W. Zhang, "Cross-modality deep feature matching network for visible-infrared person re-identification," in *Proc. IEEE 2nd Int. Conf. Comput., Commun., Percep. Quantum Technol.*, 2023, pp. 136–140.

[13] D. Hendrycks and T. G. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations,"in *Proc. Int. Conf. Learn. Representations,* 2019, pp. 1–16.

[14] D. Hendrycks, N. Mu, E. D. Cubuk, B. Zoph, J. Gilmer, and B. Lakshminarayanan, "AugMix: A simple data processing method to improve robustness and uncertainty," in *Proc. Int. Conf. Learn. Representations*, 2020, pp 1–15.

[15] Y. Qian and S.-K. Tang, "Pose attention-guided paired-images generation for visible-infrared person re-identification," *IEEE Signal Process. Lett.*, vol. 31, pp. 346–350, 2024.

[16] X. Qian et al., "Pose-normalized image generation for person re-identification," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 650–667.

[17] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1116–1124.

[18] W. Liu, Y. Wen, Z. Yu, and M. Yang, "Large-margin softmax loss for convolutional neural networks," in *Proc. Int. Conf. Learn. Representations*, 2016, pp. 507–516.

[19] D. Fu et al., "Unsupervised pre-training for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 14750–14759.

[20] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 779–788.

[21] D. Kinga and B. J. Adam, "A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, 2015, pp. 1–13.

[22] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.

[23] D. Fu et al., "Large-scale pre-training for person re-identification with noisy labels," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 2476–2486.

[24] Z. Ye, C. Hong, Z. Zeng and W. Zhuang, "Self-supervised person re-identification with channel-wise transformer," in *Proc. IEEE Int. Conf. Big Data*, Osaka, Japan, 2022, pp. 4210–4217.

[25] S. Xiang, D. Qian, J. Gao, Z. Zhang, T. Liu, and Y. Fu, "Rethinking person re-identification via semantic-based pretraining," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 20, no. 3, pp. 1–17, Dec. 2023.