Enhancing Diverse Intra-identity Representation for Visible-Infrared Person Re-Identification

Sejun Kim*, Soonyong Gwon*, Kisung Seo†
Seokyeong University, Seoul, Korea
{kimsejun5,gwonsy2,ksseo}@skuniv.ac.kr

1. Examples of intra-instance variance

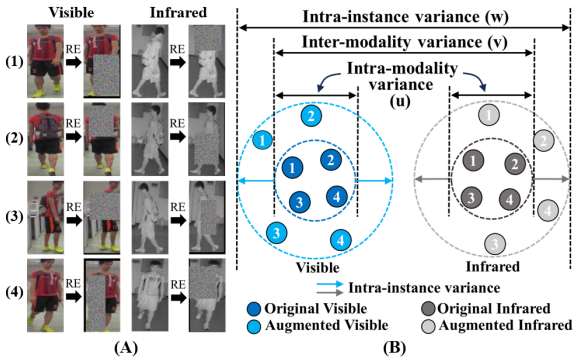


Figure 1. Example of increased modality-gap by intra-instance variance. (A) shows eight sample pairs including visible and infrared images of the same identity and applied Random Erasing (RE) [5]. (B) illustrates the feature distribution according to data augmentation for each modality. Specifically, (u) represents the intra-modality variance, while (v) is the inter-modality variance. Lastly, (w) depicts the intra-instance variance due to the utilization of data augmentation.

In this section, we describe the intra-instance variance using examples. In the case of single modality re-identification, the matching is conducted among images in the first column of Figure 1-(A). Therefore, only the intra-modality variance (u) in Figure 1-(B) needs to be addressed. On the other hand, in cross-modality re-identification, the matching is performed between the visible and infrared images in the first and third columns of Figure 1-(A) respectively. As shown in Figure 1-(B), the difference between Visible and Infrared images can cause inter-modality variance (v), which must also be considered. Existing methods conduct data augmentation, such as Random Erasing [5], to improve matching performance, as shown in the second and fourth columns of Figure 1-(A).

However, there is a limit to reduce the modality gap because it makes matching more difficult due to variations in features. As a result, it is trained to be more focused on dominant feature extraction, as shown Figure 1 in main paper. In other words, intra-instance variance (w) is necessary to be considered for the increased variance caused by the growing discrepancy from the representative features.

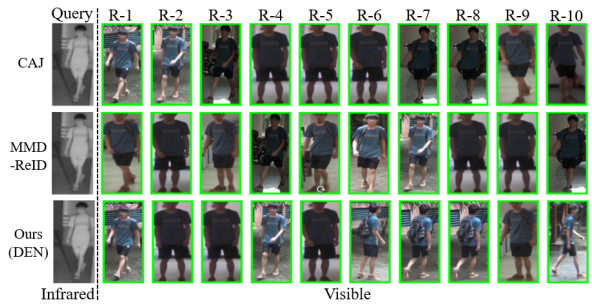
To effectively learn the features of samples with increased variance, flexible expansion of the representation space is required to enable diverse feature representations. Our method for extending Intra-modality Intra-identity Representation Space (IIRS) successfully learns diverse features through data augmentation, resulting in a more discriminative representation.

2. Retrieval Results

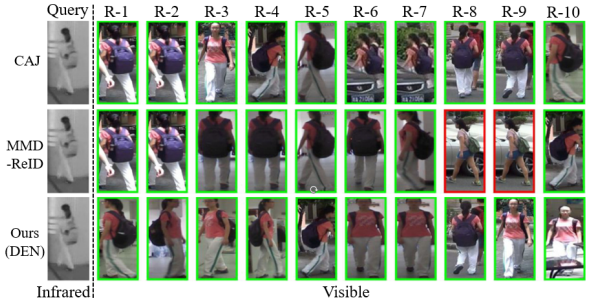
To verify diverse feature representations capability of DEN (Ours), we visualize retrieval results of easy and hard samples on SYSU-MM01 dataset. In addition, the proposed method is compared with CAJ [3], which employs channel augmentation for sample diversity, and metric learning-based MMD-ReID [2] for diversifying features.

Easy sample Retrieval. Figure 2 shows the top-10 retrieval results for easy samples with correct matching, both for existing methods and our proposed method. Existing methods tend to rank only those gallery images that closely resemble the query image at the top. In contrast, our proposed method achieves a balanced ranking of various scenes (front, side, and back) of an individual. This demonstrates its robustness to variations such as pose, camera view, and lighting, as well as its ability to utilize diverse features. Specifically, in Figure 2a, where the query image features a person's front view, existing methods tend to extract dominant features related to the front view and consequently retrieve images from the gallery that closely resemble this view. In contrast, our proposed method not only matches front views but also back views, demonstrating its ability to extract diverse features. In Figure 2b, when the query image features a person's back view with a bag, existing

*Co-first Author, equally contribute
†Corresponding Author
This work was supported by National Research Foundation of Korea Grant RS-2023-00244355 funded by the Korea government.

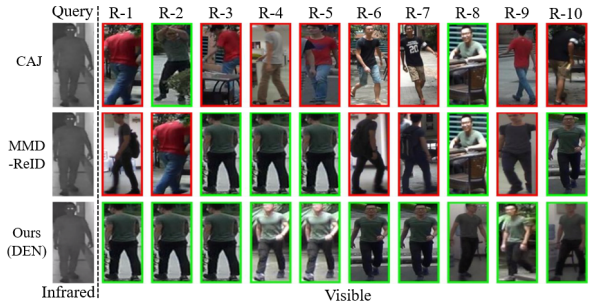


(a) Example results for an easy query sample having a front view.

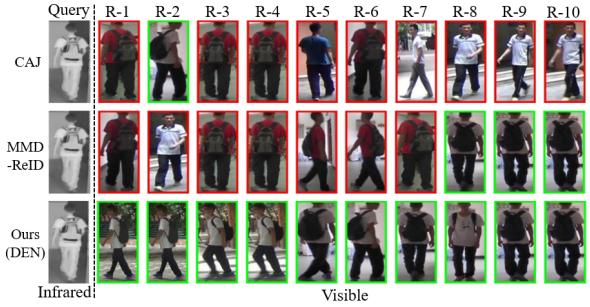


(b) Example results for a recognizable query sample having back view.

Figure 2. Comparison of retrieval results between our method and existing ones for easy samples.



(a) Example results for a unrecognizable query sample.



(b) Example results for a hard query sample having dominant feature.

Figure 3. Comparison of retrieval results between our method and existing ones for hard samples.

methods prioritize features associated with the back view and retrieve similar back views from the gallery. However, our proposed method also effectively matches front views. Therefore, our approach's enhanced recognition performance can be attributed to its ability to extract diverse features beyond dominant ones.

Hard sample Retrieval. As shown in Figure 3, the proposed method shows better retrieval results than the existing methods for hard samples. The reason for this is that the proposed method can represent various features by metric learning with considering intra-instance variance. In Figure 3a, existing methods tend to focus on incorrect dominant features, leading to poor retrieval results. In contrast, our method performs well by correctly representing various features. Similarly, in Figure 3b, existing methods emphasize dominant features, such as a bag, and incorrectly match the back of a person wearing the bag in the top rankings. In contrast, our proposed method correctly matches various images, including the front, back, and side views of the corresponding identity.

3. Implementation details

Following [2–4], we adopted an ImageNet pre-trained ResNet50 [1] as the backbone network, using only the first convolutional layer for extracting modality-specific features, while sharing the remaining parts. We apply resizing to 3 * 288 * 144, as well as random horizontal flip and random erasing to input images during train. The initial learning rate is set to $1 * 10^{-2}$ for warm-up and then increases to $1 * 10^{-1}$ after 10 epochs. The learning rate is divided by 10 at 30 epochs and 60 epochs, and training continues for a total of 100 epochs. The SGD optimizer is used for training, with the momentum and weight decay set to 0.9 and $1 * 10^{-4}$, respectively. We use the PK sampler (P identity, K sample), where each mini-batch consists of 8 identities, with each identity composed of 4 Visible images, 4 Infrared images, and 4 HueGray augmented images. The total mini-batch size is set to 96. The hyperparameters of loss factor λ_{ID}, λ_{HT}, λ_{IRD} are set to 1.0, 0.5, 0.25, and hyperparameters of margin $m_{PE}^{V,I}$, $m_{PE}^{HG,I}$, $m_{NE}^{V,I}$, $m_{NE}^{HG,I}$, m_{HT} are set to 0.1, 0.1, 0.5, 0.5, 0.5, respectively. Detailed hyperparameters setting are addressed in section 4. Our model is trained on a single Nvidia RTX 3090 GPU. Total training time is 7.5h on SYSU-MM01 dataset, and 2.1h on RegDB dataset. Test time and number of parameters for inference are same as AGW [4], which usually used base model in VI-ReID.

4. Description of hyperparameters Setting

In this section, we introduce the hyperparameters setting of loss factor (λ), margin (m). Basically we set the hyper-

parameters intuitively, except margin of Hardest Triplet loss (m_{HT}).

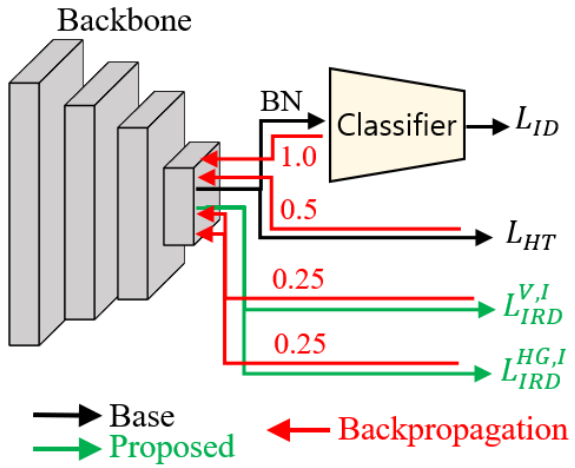


Figure 4. Visualization of the 1:1 ratio of backpropagation between the ID loss and other losses.

Loss factors. We maintain the same backpropagation ratio as the base model AGW [4] in VI-ReID for stable training. Specifically, AGW consists of identity loss (L_{ID}) and Weighted Regularization Triplet loss (L_{WRT}), and we set the backpropagation ratio for each loss to 1:1. Similarly, without additional hyperparameter experiments, we set the identity loss to 1.0 and the sum of the other loss factors to 1.0, as shown in Figure 4. However, as demonstrated in Section 3.1 and 4.4 of the main paper, the WRT of the original model interferes with learning various features. Therefore, we replace it with HT.

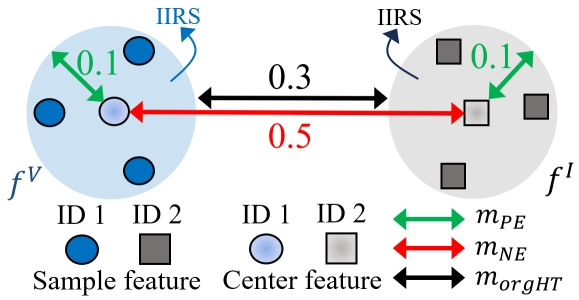


Figure 5. Visualization the margin of Visible features f^V and Infrared features f^I in embedding space. m_{orgHT} denote margin of original Hardest Triplet loss.

margin. The components of the proposed method, Intra-identity Representation Diversification (IRD), Positive Enhancement (PE), and Negative Enhancement (NE), each have their respective margins m_{PE} and m_{NE}. As shown in Figure 5, maintains a margin m_{PE} of 0.1 for IIRS, while maintains a margin m_{NE} of 0.5 for the distance between inter-identity center features. With these margin settings, it is possible to set the value of m_{HT} , which is commonly used in the original Hardest Triplet loss, to 0.3. Additionally, finding the optimal hyperparameters for margins can be a time-consuming task. Therefore, maintaining values that are commonly used can lead to faster and more stable performance, as described above.

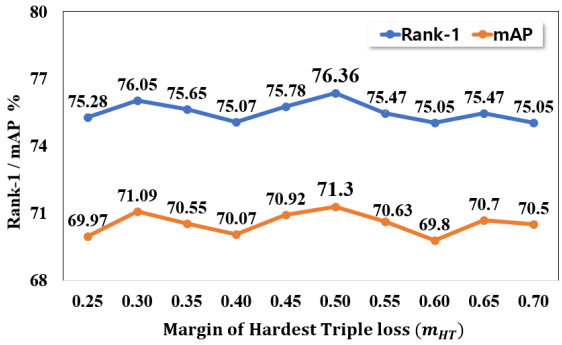


Figure 6. Influence of different m_{HT} values on our DEN.

However, we conducted hyper-parameter experiment only for the component m_{HT} of DEN. This is because, unlike IRD, which performs loss between center features, HT performs loss between sample features, making it more sensitive. As shown in Figure 6, when was varied at intervals of 0.05, m_{HT} showed the best performance at 0.5, and it also showed decent performance at 0.3.

5. HueGray for increasing intra-instance variance

In this section, we provide an example of the Hue transform and Gray transform (HG) method, which effectively generates diverse samples in the Diversity Enhancement Network (DEN) proposed in the main paper, section 3.3.

In Figure 7, row 1 illustrates that the Hue transform generates images with various colors from the original images for sample diversity. Row 2 displays images transformed into hue and gray, which resemble infrared images and increase the intra-instance variance. Consequently, HG plays an auxiliary role in reducing the modality gap by utilizing color and diversified gray images for learning, while preserving significant shape information, thus enabling discriminative learning. The ablation study in main paper verifies that even when only HG is employed, there is a significant performance improvement. Therefore, the ability to learn various features enhances retrievals for tasks characterized by large modality gaps. By combining this increased intra-instance variance with the proposed Intra-identity Representation Diversification (IRD) loss, we strive to learn diverse representations of intra-instance variance as extensively as possible, effectively reducing the modality gap and achieving superior results compared to existing methods.

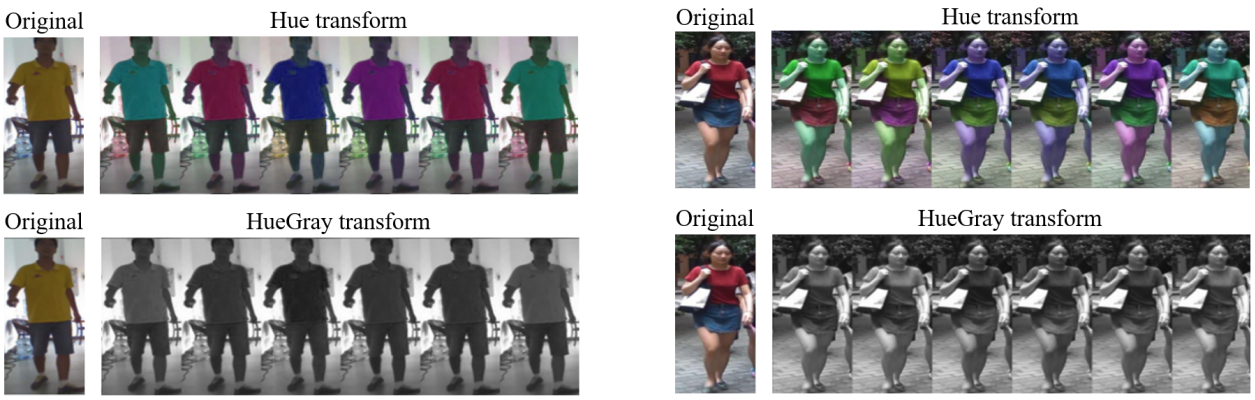


Figure 7. Illustration of the proposed Hue transform and HueGray transform.

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[2] Chaitra Jambigi, Ruchit Rawal, and Anirban Chakraborty. Mmd-reid: A simple but effective solution for visible-thermal person reid. In *British Machine Vision Conference*, 2021. 1, 2

[3] Mang Ye, Weijian Ruan, Bo Du, and Mike Zheng Shou. Channel augmented joint learning for visible-infrared recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13567–13576, 2021. 1, 2

[4] 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*, 44(6):2872–2893, 2021. 2, 3

[5] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, pages 13001–13008, 2020. 1