PartMix: Regularization Strategy to Learn Part Discovery for Visible-Infrared Person Re-identification -Supplementary Materials-

Minsu Kim[1], Seungryong Kim[2], Jungin Park[1], Seongheon Park[1], Kwanghoon Sohn[1,3*]
[1] Yonsei University [2] Korea University [3] Korea Institute of Science and Technology (KIST)
{minsukim320, newrun, sam121796, khsohn}@yonsei.ac.kr seungryong_kim@korea.ac.kr

In this supplementary material, we provide additional experimental results, implementation details, and qualitative results to complement the main paper.

A. t-SNE Visualization

Visualization for different identities images. To explain the effectiveness of our PartMix, we show the feature distribution of part descriptor with different identities in Fig. 1. For visualizing the feature distribution, the complex feature distributions are transformed into two-dimensional points based on t-SNE [2]. Each color represents the M different part maps. We can confirm that t-SNE visualization of part descriptors that have different semantic meanings are clustered into distinct groups. And we can also find that the part descriptor with the same human part information (*e.g.*, short sleeve) are clustered into the same groups. In Fig. 2, we visualize an additional example for the feature distribution of part descriptors with different identities. These two images do not share the human parts information, and thus our PartMix effectively divides part descriptors into different groups. By this visualization, we can demonstrate that our PartMix can capture different human part information and synthesize unseen combination of human parts (*i.e.* the unseen identity), improving generalization ability on unseen identity as demonstrated in the Sec 4.4 of the main paper. In addition, it can distinguish the different person identities through the combination of human parts.

B. Loss functions

Following the baseline [4], we adopt several losses, including modality learning loss $\mathcal{L}_{\text{ML}}$, modality specific ID loss $\mathcal{L}_{\text{sid}}$, center cluster loss $\mathcal{L}_{\text{cc}}$, and identity classification loss $\mathcal{L}_{\text{id}}$. In this section, we describe these losses in detail.

Modality Learning Loss. Modality learning loss [4] aims to encourage the modality-specific classifier to estimate consistent classification scores for the same identity features regardless of the modality. We make the classification scores of visible (infrared) person descriptors estimated by the visible (infrared) and mean infrared (visible) specific classifier to be similar through the KL divergence, and thus the model learns modality invariant person descriptors.

$$\begin{aligned}\mathcal{L}_{\text{ML}} = &\sum_{w=1}^{N_v} d_{KL}(\mathcal{C}_v(d_w^v)||\tilde{\mathcal{C}}_r(d_w^v)) \\ &+ \sum_{q=1}^{N_r} d_{KL}(\mathcal{C}_r(d_q^r)||\tilde{\mathcal{C}}_v(d_q^r)),\end{aligned} \tag{1}$$

where $\mathcal{C}_v(\cdot)$, $\mathcal{C}_r(\cdot)$ denote visible and infrared classifiers, and the mean classifiers of those ones are $\tilde{\mathcal{C}}_v(\cdot)$, $\tilde{\mathcal{C}}_r(\cdot)$, respectively.

Modality Specific ID Loss. For modality learning loss, we train the modality-specific classifiers to learn modality-specific knowledge from visible and infrared person descriptors such that

$$\begin{aligned}\mathcal{L}_{\text{sid}} = &-\frac{1}{N_v}\sum_{w=1}^{N_v} y_w^v \log(\mathcal{C}_v(d_w^v)) \\ &- \frac{1}{N_r}\sum_{q=1}^{N_r} y_q^r \log(\mathcal{C}_r(d_q^r)),\end{aligned} \tag{2}$$

where $\mathcal{C}_v$ and $\mathcal{C}_r$ are visible and infrared classifier.

Center Cluster Loss. To enhance the discriminative power of the person descriptor, we adopt center cluster loss [4] to penalize the distances between the person descriptors and their corresponding identity centers.

$$\begin{aligned}\mathcal{L}_{\text{cc}} = &\frac{1}{N}\sum_{i=1}^{N} ||f_i^t - z_{y_i}||_2 \\ &+ \frac{2}{P(P-1)}\sum_{k=1}^{P-1}\sum_{d=k+1}^{P} [\rho - ||z_{y_k} - z_{y_d}||_2]_+,\end{aligned} \tag{3}$$

*Corresponding author

where z_{y_i},z_{y_k}, and z_{y_d} is the mean descriptor that correspond to the y_i, y_k, and y_d identity in mini-batch, P is the number of identity in the mini-batch, and ρ is the least margin between the centers.

ID Loss To learn identity-specific feature representation across the modalities, we adopt cross-entropy loss between the identity probabilities and their ground-truth identities as follows:

$$\mathcal{L}_{\text{id}} = -\frac{1}{N^v}\sum_{i=1}^{N^v} y_i^v \log(\mathcal{C}(d_i^v)) - \frac{1}{N^r}\sum_{i=1}^{N^r} y_i^r \log(\mathcal{C}(d_i^r)), \tag{4}$$

where $\mathcal{C}(\cdot)$ is an identity classifier.

C. More Implementation Details

Training Details. To train our network, we first conduct warm up the baseline [4] for 20 epochs, to stabilize the part detector at the early stage of training and boost the convergence of training. For a fair comparison with the baseline [4], we then optimize the model for 100 epochs using overall losses. We also adopt random cropping, random horizontal flipping, and random erasing [7] for data augmentation. We set 128 images for each mini-batch. For each mini-batch, we randomly sample 8 images with 16 identities and the images are re-sized as 384×128. We select positive samples and negative samples through the entropy-based mining module. For each training sample, we set the number of positive U' and negative samples Q' as 2 and 20. To optimize the model, we utilize the Adam optimizer, where the initial learning rate is set to 3.5×10^{-4}, which decays at 80^{th} and 120^{th} epoch with a decay factor of 0.1. Through the cross-validation using grid-search, we set the hyper-parameters λ_{sid}, λ_{ML}, λ_{cont}, and λ_{aid} as 0.5, 2.5, 0.5, and 0.5, respectively. The proposed method was implemented in the Pytorch library [1]. We conduct all experiments using a single RTX A6000 GPU.

D. Other Regularization Methods Details

Mixup [6]. Following the work [6], we synthesize the mixed images by linearly interpolating image and label pairs such that

$$\begin{aligned} \tilde{x} &= \lambda x^1 + (1-\lambda)x^2, \\ \tilde{y} &= \lambda y^1 + (1-\lambda)y^2, \end{aligned} \tag{5}$$

where x^1, x^2 are randomly sampled images in mini-batch regardless of their modality, y^1, y^2 are its corresponding identity, and λ is the combination ratio sampled from the beta distribution $Beta(\alpha, \alpha)$, where the α is set to 1.

Manifold MixUp [3]. We also synthesize the mixed training samples using Manifold MixUp [3] that applies MixUp [6] in the hidden feature space as follows :

$$\begin{aligned} \tilde{x} &= \lambda \mathcal{E}_g(x^1) + (1-\lambda)\mathcal{E}_g(x^2), \\ \tilde{y} &= \lambda y^1 + (1-\lambda)y^2, \end{aligned} \tag{6}$$

where $\mathcal{E}_g(x)$ denotes a forward pass until randomly chosen layer g. We also sample the combination ratio λ from the beta distribution $\beta(\alpha, \alpha)$, where the α is set as 1.

CutMix [5]. We generate training samples with CutMix operation as follows:

$$\begin{aligned} \tilde{x} &= \mathbf{M} \odot x^1 + (\mathbf{1} - \mathbf{M}) \odot x^2, \\ \tilde{y} &= \lambda y^1 + (1-\lambda)y^2, \end{aligned} \tag{7}$$

where $\mathbf{M}$ is a binary mask, $\mathbf{1}$ is a binary mask filled with ones, $\odot$ is element-wise multiplication, and the setting of λ is identical to Mixup [6]. To sample the mask $\mathbf{M}$, we uniformly sample the bounding box coordinates $\mathbf{B} = (b_x, b_y, b_w, b_h)$ such that

$$\begin{aligned} b_x &\sim Unif(0, W), b_w = W\sqrt{1-\lambda}, \\ b_y &\sim Unif(0, H), b_h = H\sqrt{1-\lambda}, \end{aligned} \tag{8}$$

where W, H is width and height of the person image and $Unif(\cdot, \cdot)$ denotes a uniform distribution.

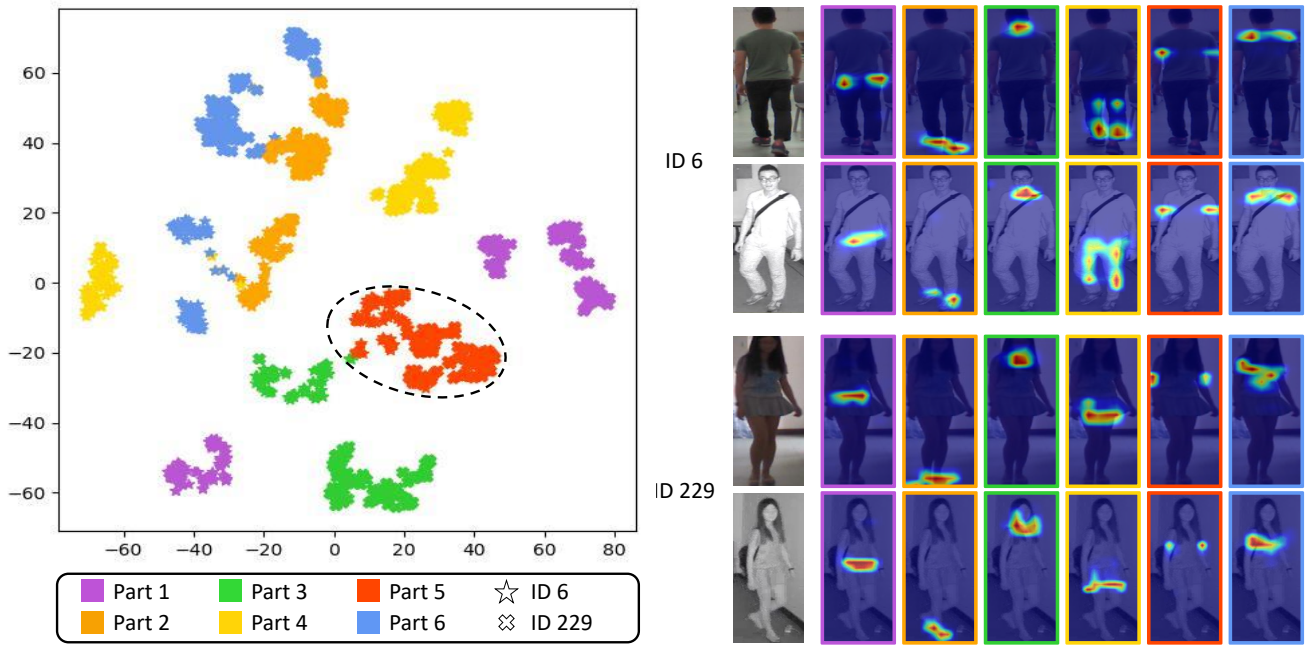

Figure 1. **Visualization on the feature distribution of part descriptor with different identity images.** Data projection in 2-D space is attained by t-SNE based on the feature representation. Each color represents the different human parts. Our PartMix effectively clusters the same human part information (*e.g.*, short sleeve) in the same group (represented using a dotted circle), while the different human parts are divided into different groups.

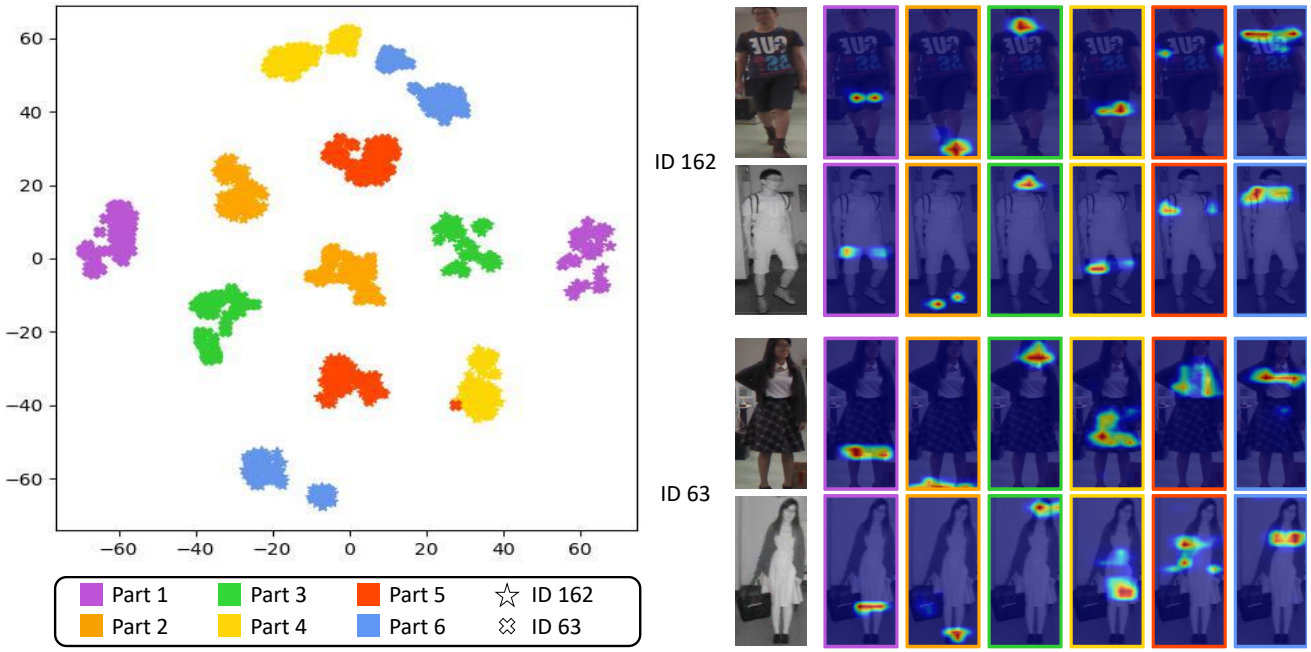

Figure 2. **Visualization of the feature distribution of part descriptor with different identity images.** The details are the same as above.

References

[1] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 2

[2] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 1

[3] Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. Manifold mixup: Better representations by interpolating hidden states. In *International Conference on Machine Learning*, pages 6438–6447. PMLR, 2019. 2

[4] 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 reidentification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4330–4339, 2021. 1, 2

[5] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6023–6032, 2019. 2

[6] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. Mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018. 2

[7] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13001–13008, 2020. 2