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

Minsu Kim¹, Seungryong Kim², Jungin Park¹, Seongheon Park¹, Kwanghoon Sohn¹,³,*
¹ Yonsei University ² Korea University ³ Korea Institute of Science and Technology (KIST)
{minsukim320, newrun, saml21796, khsohn}@yonsei.ac.kr seungryong_kim@korea.ac.kr

Abstract

Modern data augmentation using a mixture-based technique can regularize the models from overfitting to the training data in various computer vision applications, but a proper data augmentation technique tailored for the part-based Visible-Infrared person Re-IDentification (VI-ReID) models remains unexplored. In this paper, we present a novel data augmentation technique, dubbed PartMix, that synthesizes the augmented samples by mixing the part descriptors across the modalities to improve the performance of part-based VI-ReID models. Especially, we synthesize the positive and negative samples within the same and across different identities and regularize the backbone model through contrastive learning. In addition, we also present an entropy-based mining strategy to weaken the adverse impact of unreliable positive and negative samples. When incorporated into existing part-based VI-ReID model, PartMix consistently boosts the performance. We conduct experiments to demonstrate the effectiveness of our PartMix over the existing VI-ReID methods and provide ablation studies.

1. Introduction

Person Re-IDentification (ReID), aiming to match person images in a query set to ones in a gallery set captured by non-overlapping cameras, has recently received substantial attention in numerous computer vision applications, including video surveillance, security, and persons analysis [64, 76]. Many ReID approaches [1, 2, 25, 26, 32, 41, 59, 73, 78] formulate the task as a visible-modality retrieval problem, which may fail to achieve satisfactory results under poor illumination conditions. To address this, most surveillance systems use an infrared camera that can capture the scene even in low-light conditions. However, directly matching these infrared images to visible ones for ReID poses additional challenges due to an inter-modality variation [60, 62, 65].

To alleviate these inherent challenges, Visible-Infrared person Re-IDentification (VI-ReID) [5–7, 12, 24, 35, 48, 54, 60–62, 66] has been popularly proposed to handle the large intra- and inter-modality variations between visible images and their infrared counterparts. Formally, these approaches first extract a person representation from whole

*Corresponding author

This research was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (NRF2021R1A2C2006703).
visible and infrared images, respectively, and then learn a modality-invariant feature representation using feature alignment techniques, e.g., triplet [5,7,24,60–62] or ranking criterion [12,66], so as to remove the inter-modality variation. However, these global feature representations solely focus on the most discriminative part while ignoring the diverse parts which are helpful to distinguish the person identity [53,63].

Recent approaches [53,57,63] attempted to further enhance the discriminative power of person representation for VI-ReID by capturing diverse human body parts across different modalities. Typically, they first capture several human parts through, e.g., horizontal stripes [63], clustering [53], or attention mechanisms [57] from both visible and infrared images, extract the features from these human parts, and then reduce inter-modality variation in a part-level feature representation. Although these methods reduce inter-modality variation through the final prediction (e.g., identity probability), learning such part detector still leads to overfitting to the specific part because the model mainly focuses on the most discriminative part to classify the identity, as demonstrated in [4,15,31,52]. In addition, these parts are different depending on the modality, it accumulates errors in the subsequent inter-modality alignment process, which hinders the generalization ability on unseen identity in test set.

On the other hand, many data augmentation [19,22,38,45,68,69] enlarge the training set through the image mixture technique [69]. They typically exploit the samples that linearly interpolate the global [38,44,69] or local [19,68] images and label pairs for training, allowing the model to have smoother decision boundaries that reduce overfitting to the training samples. This framework also can be a promising solution to reduce inter-modality variation by mixing the different modality samples to mitigate overfitting to the specific modality, but directly applying these techniques to part-based VI-ReID models is challenging in that they inherit the limitation of global and local image mixture methods (e.g., ambiguous and unnatural patterns, and local patches with only background or single human part). Therefore, the performance of part-based VI-ReID with these existing augmentations would be degraded.

In this paper, we propose a novel data augmentation technique for VI-ReID task, called PartMix, that synthesizes the part-aware augmented samples by mixing the part descriptors. Based on the observation that learning with the unseen combination of human parts may help better regularize the VI-ReID model, we randomly mix the inter- and intra-modality part descriptors to generate positive and negative samples within the same and across different identities, and regularize the model through the contrastive learning. In addition, we also present an entropy-based mining strategy to weaken the adverse impact of unreliable positive and negative samples. We demonstrate the effectiveness of our method on several benchmarks [33,55]. We also provide an extensive ablation study to validate and analyze components in our model.

2. Related work

Person ReID. Person Re-IDentification (ReID) aims to search a target person from a large gallery set, where the images are captured from non-overlapping visible camera views. With the advent of deep convolutional neural networks (CNNs), to solve this task, existing works [1,3,46] encourage the person representation within the same identity to be similar through feature-level constraint, including triplet constraint [46], quadruplet constraint [3], or group consistency constraint [1]. However, since these methods learn features from the whole person image, they often suffer from intra-modality variation caused by the human pose variation and part occlusions [41]. Thus, there have been many efforts to focus on extracting human body parts that can provide fine-grained person image descriptions through uniform partitions [8,41,42,47,73] or attention mechanisms [25,28,29,59,74,77].

Visible-Infrared Person ReID. Most surveillance systems deploy infrared images to achieve satisfactory results under poor illumination conditions [20,55]. However, directly applying person re-id methods suffer from the different distribution between modalities, typically known as modality discrepancy [60,62,65]. To alleviate this, Visible-Infrared person re-id (VI-ReID) has been popularly studied to match a person between visible and infrared images with challenges posed by large intra-modality variation and inter-modality variations. Wu et al. [55] firstly introduced a large-scale VI-ReID dataset, named SYSU-MM01, and proposed a deep zero-padding strategy to explore modality-specific structure in a one-stream network. Recently, modality-invariant feature learning based VI-ReID has been proposed to project the features from different modalities into the same feature space. Formally, these methods extract features from visible and infrared person images and then reduce the inter-modal discrepancy by feature-level constraints, such as triplet constraint [7,60–62], or ranking constraint [12,66]. However, these methods usually focus only on the most discriminative part rather than the diverse parts which are helpful to distinguish different persons [53,63]. Therefore, several methods have been proposed to align inter-modality discrepancy in a fine-grained manner, exploiting horizontal stripes [63], or inter-modality nuances [53,57]. However, the part detector can be easily overfitted to the specific part. Moreover, these part maps are not semantically consistent across modalities, showing limited performance.
3. Proposed Method

3.1. Preliminaries and Problem Formulation

Let us denote visible and infrared image sets as \(\mathcal{X}^v = \{x_i^v\}_{i=1}^{N^v}\) and \(\mathcal{X}^r = \{x_i^r\}_{i=1}^{N^r}\), where \(x_i^v\) and \(x_i^r\) are images, and \(N^v\) and \(N^r\) are the number of visible and infrared images, and they are unpaired. For visible and infrared image sets, the corresponding identity label sets are defined such that \(\mathcal{Y}^v = \{y_i^v\}_{i=1}^{N^v}\) and \(\mathcal{Y}^r = \{y_i^r\}_{i=1}^{N^r}\), whose label candidates are shared. The objective of VI-ReID is to learn modality-invariant person descriptors, denoted by \(d^v\) and \(d^r\), for matching persons observed from visible and infrared cameras. For simplicity, we denote visible and infrared modalities as \(t \in \{v, r\}\) unless stated.

Most recent state-of-the-art methods are based on part-based person representation [53, 57, 63] that aims to extract discriminative human parts information and use them to enhance the robustness of person representation against human pose variation across the modality. These methods typically involve discovering diverse and discriminative human parts in an attention mechanism, and generating the person descriptor by assembling a global descriptor and part descriptors for retrieving visible (or infrared) images according to the given infrared (or visible) images.

Specifically, given visible and infrared images, the feature map for each modality is computed through an embedding network \(\mathcal{E}(\cdot)\) such that \(f^t = \mathcal{E}(x^t)\). The part detector \(\mathcal{D}(\cdot)\) then produces human parts, followed by sigmoid function \(\sigma(\cdot)\), to output part map probability, denoted by \(\{m^t(k)\}_{k=1}^{M^t} = \sigma(\mathcal{D}(f^t))\), where \(M^t\) is the number of part maps. The part descriptors are then formulated as follows:

\[
p^t = [p^t(k)]_{k=1}^{M^t} = [\text{GAP}(m^t(k) \odot f^t)]_{k=1}^{M^t},
\]

where \(\text{GAP}(\cdot)\) denotes a global average pooling, \(\odot\) is an element-wise multiplication, and \([\cdot]\) is a concatenate operation. Note that they apply element-wise multiplication between \(m^t(k)\) and each channel dimension in \(f^t\). They finally concatenate the global descriptor \(g^t\) such that \(l^t = \text{GAP}(f^t)\) and part descriptors \(p^t\) to obtain person descriptor \(d^t\) for matching the persons observed from visible and
infrared cameras such that

\[ d^t = [g^t, p^t]. \]  

To train such a model, since only identity labels are available, they adopted a cross-entropy loss between the identity probabilities and ground-truth identities. In addition, they also adopted several loss functions, including knowledge distillation [57] or metric learning [53, 57, 63] loss, to learn modality invariance in a part-level feature representation.

While these losses let the network focus on human parts across modalities and reduce inter-modality variations by aligning person descriptors within the same identity, the part detectors learned by these methods have been often overfitted to the specific part, as exemplified in Fig. 4. In addition, these learned parts may not be semantically consistent across modalities [14, 17, 23]. Therefore, they fail to localize discriminative and semantically-consistent human parts across the visual and infrared modalities, thus showing limited performance.

### 3.2. Overview

To overcome the limitations of previous part-based person representation methods [53, 57, 63], our PartMix, which we introduce in the remainder of this section, accounts for the observation that learning with the part-aware augmented samples across both inter- and intra-modality can help better regularize the model. Unlike conventional methods [19, 38, 44, 68, 69] that exploit image mixture techniques that yield unnatural patterns, and local patches with only background or single human part, we present a novel part mixing strategy to synthesize augmented samples by mixing partial descriptors across the modalities, and use them to synthesize positive and negative samples to maximize the similarities of positive pairs and minimize the similarities of negative pairs through the contrastive learning objective. It helps to regularize the model and mitigate the overfitting to the specific part and modality, improving the generalization capability of the model. Furthermore, to eliminate the unreliable positive and negative samples, we present an entropy-based mining strategy that can help guide a representation to be more discriminative.

### 3.3. Part Mixing for Data Augmentation

One of the most straightforward ways of leveraging regularization to better learn part discovery may be to utilize existing augmentation techniques, e.g., using global [38, 44, 69] or local [19, 68] image mixtures. These existing strategies, however, are difficult to be directly applied to part-based VI-ReID methods [53, 57, 63] due to the following two respects. Firstly, directly applying the global image mixture methods [38, 44, 69] suffers from the locally ambiguous and unnatural pattern, and thus mixed sample confuses the model, especially for localization as demonstrated in [43, 68]. Secondly, mixing the local image region [19, 68] without part annotation may contain only a background or single human part, and thus it may cause performance degradation for part-based models which require diverse and discriminative discovery of human parts for distinguishing the person identity.

To overcome these, we present a novel data augmentation technique tailored to part-based methods, called PartMix, that mixes the part descriptors extracted from different person images. By mixing the part descriptors rather than the images, we can synthesize the augmented samples with diverse combinations of human parts. Concretely, we first collect the part descriptors in the visible and infrared modalities of the mini-batch, denoted as the descriptor bank \( P^t = \{ p_1^t, p_2^t, ..., p_N^t \} \). We then mix the part descriptors through part mix operation across the inter-modality \( A(p_i^v(u), p_j^h(h)) \) and intra-modality \( A(p_i^v(u), p_j^v(h)) \) sequentially as follows:

\[
\begin{align*}
A(p_i^v(u), p_j^h(h)) &= [p_1^v(u), ..., p_i^v(u - 1), p_j^h(h), p_i^v(u + 1), ..., p_M^v(u)], \\
A(p_i^v(u), p_j^v(h)) &= [p_1^v(u), ..., p_i^v(u - 1), p_j^v(h), p_i^v(u + 1), ..., p_M^v(u)].
\end{align*}
\]

where \( h, u \) denote the randomly sampled indexes of part descriptors \( p^t \). Note that we exclude the global descriptor \( g^t \) in the part mixing above because it contains all human body parts information.

### 3.4. Sample Generation for Contrastive Learning

Existing image mixture-based methods [19, 38, 44, 68, 69] generate the training samples by linearly interpolating the images and the corresponding labels. These approaches, however, only synthesize the samples with the combination of identities in the training set, and thus they have limited generalization ability on the VI-ReID task where identities...
in the test set are different from the training set. To alleviate this, we present a sample generation strategy that can synthesize positive and negative samples with the unseen combination of human parts (i.e., the unseen identity). In the following section, we explain how to achieve positive bank \( B_{i}^{+,\tau} \) and negative bank \( B_{i}^{-,\tau} \) in detail. For simplicity, only visible samples are described as an example.

**Positive Samples.** Our first insight is that the combination of the human parts of the persons with the same identity has to be consistent. To this end, we design positive samples that mix the same part information between the person images within the same identity. Specifically, we mix the part descriptors with the same identity using (3). Each positive sample for visible modality is denoted as

\[
b_{i}^{+,v} = [A(p_{i}^{v}(k),p_{j}^{v}(k)),A(p_{i}^{v}(k),p_{j}^{v}(k))], \quad \text{if } y_{i} = y_{j}.
\]

(4)

Note that we only mix the part-level descriptor within the same identity (i.e., \( y_{i} = y_{j} \)).

**Negative Sample.** The positive samples in (4) encourage the part-level feature within the same identity across the inter- and intra-modality to be invariant. This, however, does not guarantee that the model can distinguish the person identity with different combinations of human parts, and localize the diverse human parts within the same person.

To overcome this, we design the negative samples that encourage the model to distinguish the person identity when the combination of human parts is different and localize diverse human parts in each identity. We mix the part descriptors within the same and different identity using (3) as

\[
b_{i}^{-,v} = \begin{cases} [A(p_{i}^{v}(k),p_{j}^{v}(h)),A(p_{i}^{v}(k),p_{j}^{v}(h))], \quad \text{if } y_{i} = y_{j} \\ [A(p_{i}^{v}(k),p_{j}^{v}(k)),A(p_{i}^{v}(k),p_{j}^{v}(k))], \quad \text{if } y_{i} \neq y_{j} \end{cases}
\]

(5)

where \( k \) and \( h \) denote the different indexes of part descriptors. Note that our negative samples cover the unseen combination of human parts in the training set. Therefore, these samples can be seen as out-of-distribution negative samples that can provide supportive information for improving the generalization capability of the model as investigated in [9,40].

3.5. Entropy-based Mining Strategy

Even though the proposed sample generation strategy through part mixing yields comparable performance to some extent (which will be discussed in experiments), it relies on the assumption that there are different human part information (e.g., clothes, hairstyle) for each person. However, several persons with different identities share a similar appearance (e.g., wearing similar clothes). Therefore, simply mixing these samples makes false negative samples that still have the same combination of human parts. Moreover, the false positive samples can be synthesized by randomly mixing human parts that have different semantic meanings, especially in the early training stage.

To overcome this, inspired by the uncertainty-based approaches [10,37], we present an entropy-based mining strategy for eliminating the false positive and false negative samples. We exploit the difference between the pairwise entropy of the identity prediction that can be an uncertainty measure for positive and negative samples, where the reliable samples are determined as a positive pair with a smaller entropy difference and a negative pair with a larger entropy difference. We first obtain identity probability for each sample through the part-level identity classifier \( C_{p}(\cdot) \). We then compute the difference between the pairwise entropy for positive and negative samples as follows:

\[
\begin{align*}
\hat{h}_{i}^{+,v} &= \frac{[H(C_{p}(p_{i}^{v})) - H(C_{p}(b_{i}^{+,v}(j)))]}{U}, \\
\hat{h}_{i}^{-,v} &= \frac{[H(C_{p}(p_{i}^{v})) - H(C_{p}(b_{i}^{-,v}(j)))]}{Q},
\end{align*}
\]

(6)

where \( H(\cdot) \) is entropy function [37], and \( U, Q \) are the number of positive and negative pairs. These uncertainty value sets for positive \( h_{i}^{+,v} \) and negative \( h_{i}^{-,v} \) are then sorted in ascending and descending order, respectively. We select the top \( U' \) and \( Q' \) samples for positive bank \( B_{i}^{+,v} = [b_{i}^{+,v}(j)]_{j=1}^{U'} \) and negative bank \( B_{i}^{-,v} = [b_{i}^{-,v}(j)]_{j=1}^{Q'} \), respectively.

3.6. Loss Functions

In this section, we describe several loss functions to train our network. We newly present contrastive regularization loss \( L_{\text{cont}} \) and part ID loss \( L_{\text{aid}} \) to regularize the model through positive and negative samples.

**Contrastive Regularization Loss.** We aim to maximize the similarities between positive pairs and minimize similarities between negative pairs. To this end, inspired by [34,56], we adopt the contrastive regularization loss that jointly leverages positive and negative samples to regularize the model, and thus the model can overcome the limitation of strong reliance on supervision signal. We first define the similarity between part descriptors and positive (negative) samples, defined as \( s_{i,j} = \text{sim}(p_{i},b_{i}^{+}(j)) \) \((s_{i,k} = \text{sim}(p_{i},b_{i}^{-}(k)))\), where \( \text{sim}(\cdot,\cdot) \) denotes a similarity function. The contrastive loss can be written as follows:

\[
L_{\text{cont}} = \sum_{i=1}^{N^{+}+N^{-}} - \log \frac{\sum_{j=1}^{U'} \exp(s_{i,j}^{+}/\tau)}{\sum_{j=1}^{U'} \exp(s_{i,j}^{+}/\tau) + \sum_{k=1}^{Q'} \exp(s_{i,k}^{-}/\tau)},
\]

(7)

where \( \tau \) is a scaling temperature parameter.

**Part ID Loss.** To estimate accurate entropy values for the augmented samples, we adopt part ID loss for an additional
part classifier that provides identity scores from part descriptors. It allows the model to eliminate the unreliable positive and negative samples, and learn discriminative and semantically consistent part discovery, simultaneously.

\[ L_{\text{aid}} = -\frac{1}{N} \sum_{i=1}^{N} y_i \log(p_i) \]. \hspace{1cm} (8)

**Total Loss.** Following the baseline [57], we also adopt modality learning loss \( L_{\text{ML}} \), modality specific ID loss \( L_{\text{sid}} \), center cluster loss \( L_{\text{cc}} \), and identity classification loss \( L_{\text{id}} \). The total loss function of our approach can be written as \( L = L_{\text{id}} + L_{\text{cc}} + \lambda_{\text{sid}} L_{\text{sid}} + \lambda_{\text{ML}} L_{\text{ML}} + \lambda_{\text{aid}} L_{\text{aid}} + \lambda_{\text{cont}} L_{\text{cont}} \), where \( \lambda_{\text{sid}}, \lambda_{\text{ML}}, \lambda_{\text{aid}}, \) and \( \lambda_{\text{cont}} \) are weights that control the importance of each loss.

### 3.7. Discussion

Most recent trends in inter-domain scenarios (e.g., domain adaptation [67, 70]) exploited counterfactual intervention to learn domain invariant knowledge, improving generalization capabilities. These frameworks consist of two main components: generating “counterfactual samples” by changing the domain (e.g., style) and using them in the model training for “intervention”. Our PartMix satisfies these conditions, as we synthesize the modality-mixed samples by changing the part descriptor across the modality and training the whole model with these samples. Therefore, we can interpret our PartMix as a counterfactual intervention for inter-modality part-discovery, where the part mixing module can be viewed as a “counterfactual” sample generator and the contrastive regularization as an “intervention”. This allows our model to learn modality invariant part representation by encouraging the part discovery to be invariant under different interventions.

### 4. Experiments

#### 4.1. Experimental Setup

In this section, we comprehensively analyze and evaluate our PartMix on several benchmarks [33, 55]. First, we analyze the effectiveness of our PartMix and comparison with other regularization methods. We then evaluate our method compared to the state-of-the-art methods for VI-ReID. In the experiment, we utilize MPANet [57] as our baseline model. Additional implementation details will be explained in the supplementary material.

**Dataset.** We evaluate our method on two benchmarks, SYSU-MM01 [55] and RegDB [33]. Firstly, SYSU-MM01 dataset [55] is a large-scale VI-ReID dataset. This dataset contains 395 identities with 22,258 visible images acquired by four cameras and 11,909 near-infrared images acquired by two cameras for the training set. The testing set contains 96 identities with 3,803 near-infrared images in the query, and 301 and 3,010 visible images in the gallery for single-shot and multi-shot, respectively. Secondly, RegDB dataset [33] contains 4,120 visible and infrared paired images with 412 person identities, where each person has 10 visible and 10 far-infrared images. Following [33], we randomly split the dataset for training and testing sets, where each set contains non-overlapping 206 person identities between the sets.

**Evaluation Protocols.** For SYSU-MM01 [55] benchmark, we follow the evaluation protocol in [55]. We test our model in all-search and indoor-search settings, where the gallery sets for the former include images captured by all four visible cameras, and the latter includes two indoor ones. For RegDB [33] benchmark, we evaluate our model on infrared to visible and visible to infrared setting, where the former retrieve infrared images from visible ones, and the latter retrieves visible ones from infrared ones. For both benchmarks, we adopt the cumulative matching character-
Table 1. Ablation study for the different components of our method on the SYSU-MM01 dataset [55].

<table>
<thead>
<tr>
<th>Methods</th>
<th>SYSU-MM01</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Single-shot</td>
<td>Multi-shot</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Rank-1  mAP</td>
<td>Rank-1  mAP</td>
<td></td>
</tr>
<tr>
<td>Base</td>
<td>70.58 68.24</td>
<td>75.58 62.91</td>
<td></td>
</tr>
<tr>
<td>Base+IntraPM</td>
<td>72.72 69.84</td>
<td>76.82 64.64</td>
<td></td>
</tr>
<tr>
<td>Base+InterPM</td>
<td>75.61 71.79</td>
<td>78.64 67.72</td>
<td></td>
</tr>
<tr>
<td>Base+IntraPM + InterPM</td>
<td>75.86 72.71</td>
<td>79.05 68.80</td>
<td></td>
</tr>
<tr>
<td>Ours</td>
<td>77.78 74.62</td>
<td>80.54 69.84</td>
<td></td>
</tr>
</tbody>
</table>

Figure 5. Ablation study for part mixing with different numbers of the mixed part (left) and number of part maps (right).

4.2. Ablation Study

In this section, we conduct ablation analyses to investigate the effectiveness of our framework. In this ablation study, all experiments are conducted on SYSU-MM01 dataset [55].

Effectiveness of Each Component. Here we analyze the key components of PartMix, including intra-modality part mixing (IntraPM), inter-modality part mixing (InterPM), and entropy-based mining (EM). For the fair comparison, we utilize MPANet [57] as our baseline (Base) for all experiments. As summarized in Table 1, each component helps to boost performance. Starting from the baseline, adding IntraPM improves the performance which indicates that IntraPM effectively mitigates intra-modality variation (e.g., human pose variation) thanks to samples synthesized by mixing part information extracted from diverse human poses. When the InterPM is added, we observe a significant improvement, which confirms that the InterPM effectively synthesizes the unseen combination of human parts across the modalities, and thus mitigates the overfitting to the specific parts and modality, simultaneously. We also note that EM also brings the performance gain for VI-ReID by mining the reliable samples among initial positive and negative samples.

Number of Mixing Parts. In our experiment, we mix the part descriptors $B$ times using (3). Note that we set the number of part maps as 6 in this experiment. We analyze the quantitative comparison with a different number of mixed parts $B$ as shown in Fig. 5. The result shows that ours with various numbers of mixed parts, which shows the effectiveness of our PartMix. We consider ours with $B = 0$ as the baseline [57]. The result with $B = 6$ is that positive and negative samples are set to the counterpart modality samples within the same and different identities, respectively. The performance has improved as $B$ is increased from 1, and after achieving the best performance at $B = 2$ the rest showed slightly improved performance. The lower performance with the larger number of mixed parts indicates that easy samples in which the combination of human parts is not significantly different from samples with a different identity in mini-batch can actually be less effective in regularization. Since the result with $B = 2$ has shown the best performance on the SYSU-MM01 dataset [55] and RegDB dataset [33], we set $B$ as 2 for all experiments.

Number of Part Maps. In Fig. 5, we evaluate our model with the different number of part maps. In all experiments, we set the number of mixed parts to 1/3 of the number of part maps. The results show that our PartMix consistently boosts the performance of mAP in single-shot all-search on the SYSU-MM01 dataset. For $M = 1$, we exploit training samples with the same and different identities in the mini-batch as positive and negative samples, respectively. We consider ours with $M = 1$ as contrastive learning with the global descriptor. It shows that contrastive learning effectively regularizes the model, and thus the model can mitigate modality discrepancy. Specifically, as $M$ is increased from 4, the performance of our model converges to high mAP. These results indicate that our PartMix can consistently capture diverse and discriminative human parts with only a small number of part maps. Since the result with $M = 6$ has shown the best performance of mAP in single-shot all-search on the SYSU-MM01 dataset, we set $M = 6$ for the remaining experiments.

4.3. Comparison to Other Regularization Methods

In this section, we validate the effectiveness of our PartMix through the comprehensive comparison with other regularization methods, including MixUp [69], Manifold MixUp [44], and CutMix [68]. Table 2 shows PartMix significantly outperforms all other regularization methods [44, 68, 69]. Interestingly, the MixUp method [69]
Table 3. Quantitative evaluation on SYSU-MM01 dataset [55] and RegDB dataset [33]. For evaluation, we measure Rank-1 accuracy(%) and mAP(%). Our results show the best results in terms of Rank-1 accuracy and mAP.

<table>
<thead>
<tr>
<th>Method</th>
<th>SYSU-MM01 [55]</th>
<th>RegDB [33]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All-Search</td>
<td>Indoor-Search</td>
</tr>
<tr>
<td></td>
<td>Single-Shot</td>
<td>Multi-Shot</td>
</tr>
<tr>
<td></td>
<td>Rank-1 mAP</td>
<td>Rank-1 mAP</td>
</tr>
<tr>
<td>Two-stream [48]</td>
<td>11.65 12.85</td>
<td>16.33 8.03</td>
</tr>
<tr>
<td>One-stream [48]</td>
<td>12.04 13.67</td>
<td>16.26 8.59</td>
</tr>
<tr>
<td>Zero-Padding [48]</td>
<td>14.80 15.95</td>
<td>19.13 10.89</td>
</tr>
<tr>
<td>cmGAN [5]</td>
<td>26.97 27.80</td>
<td>31.49 22.27</td>
</tr>
<tr>
<td>D²RL [50]</td>
<td>28.90 29.20</td>
<td>-</td>
</tr>
<tr>
<td>JSIA-ReID [49]</td>
<td>38.10 39.00</td>
<td>45.10 29.50</td>
</tr>
<tr>
<td>AlignGAN [48]</td>
<td>42.40 40.70</td>
<td>51.50 33.90</td>
</tr>
<tr>
<td>DFE [11]</td>
<td>48.71 48.59</td>
<td>54.63 42.14</td>
</tr>
<tr>
<td>XIV-ReID [24]</td>
<td>49.92 50.73</td>
<td>-</td>
</tr>
<tr>
<td>CMM+CML [27]</td>
<td>51.80 51.21</td>
<td>56.27 43.39</td>
</tr>
<tr>
<td>SIM [18]</td>
<td>56.93 60.88</td>
<td>-</td>
</tr>
<tr>
<td>CoAL [51]</td>
<td>57.22 57.20</td>
<td>-</td>
</tr>
<tr>
<td>DG-VAE [36]</td>
<td>59.49 58.46</td>
<td>-</td>
</tr>
<tr>
<td>cm-SSFT [30]</td>
<td>61.60 63.20</td>
<td>63.40 62.00</td>
</tr>
<tr>
<td>SMCL [54]</td>
<td>67.39 61.78</td>
<td>72.15 54.93</td>
</tr>
<tr>
<td>MPANet [57]</td>
<td>70.58 68.24</td>
<td>75.58 62.91</td>
</tr>
<tr>
<td>MSCLNet [72]</td>
<td>76.99 71.64</td>
<td>-</td>
</tr>
<tr>
<td>Ours</td>
<td>77.78 74.62</td>
<td>80.54 69.84</td>
</tr>
</tbody>
</table>

Highly degrades the performance for part-based VI-ReID. It demonstrates that simply applying MixUp to the part-based VI-ReID degrades the localization ability of the model due to ambiguous and unnatural patterns in mixed images, and thus the model fails to distinguish the different person identities as done in the literature [43, 68]. On the other hand, Manifold MixUp [44] shows slightly improved rank-1 accuracy, but achieves lower performance in mAP than the baseline. The result shows that it still inherits the limitation of the global mixture model [38, 44, 69]. Although CutMix [68] achieves improved performance than the baseline [57], it achieves relatively lower performance than ours by 3.93% and 5.08% mAP in single-shot all-search and multi-shot all-search on the SYSU-MM01 dataset. It demonstrated that our method effectively alleviates the overfitting to the specific part and modality in part-based VI-ReID. Based on all these evaluation and comparison results, we can confirm the effectiveness of our methods.

4.4. Comparison to Other Methods

In this section, we evaluate our framework through comparison to state-of-the-art methods for VI-ReID, including Two-stream [48], One-stream [48], Zero-Padding [48], cmGAN [5], D²RL [50], JSIA-ReID [49]. AlignGAN [48], DFE [11], XIV-ReID [24], CMM+CML [27], SIM [18], CoAL [51], DG-VAE [36], cm-SSFT [30]. SMCL [54], MPANet [57], and MSCLNet [72].

Results on SYSU-MM01 dataset. We evaluate our PartMix on SYSU-MM01 benchmark [55] as provided in Table 3. PartMix achieves the Rank-1 accuracy of 77.78% and mAP of 74.62% in all-search with single-shot mode, improving the Rank-1 accuracy by 0.79% and mAP by 2.98% over the MSCLNet [72]. In indoor-search with single-shot mode, our PartMix outperforms the MSCLNet [72] by Rank-1 accuracy of 3.03% and mAP of 3.21%.

Results on RegDB dataset. We also evaluate our method on RegDB benchmark [33]. As shown in Table 3, PartMix records state-of-the-art results with the Rank-1 accuracy of 84.93% and mAP of 82.52% in infrared to visible and the Rank-1 accuracy of 85.66% and mAP of 82.27% in visible to infrared mode. Our PartMix outperforms the Rank-1 accuracy by 1.07% and mAP by 4.21% in infrared to visible mode and the Rank-1 accuracy by 1.49% and mAP by 1.28% in visible to infrared mode over the MSCLNet [72].

5. Conclusion

In this paper, we have presented a novel data augmentation technique, called PartMix, that generates part-aware augmented samples by mixing the part descriptors. We introduce a novel sample generation method to synthesize the positive and negative samples and an entropy-based mining strategy to select reliable positive and negative samples to regularize the model through the contrastive objective. We have shown that PartMix achieves state-of-the-art performance over the existing methods on several benchmarks.

Acknowledgements. This research was supported by the Yonsei Signature Research Cluster Program of 2022 (2022-22-0002) and the KIST Institutional Program (Project No.2E32283-23-064).
References


of the IEEE/CVF conference on computer vision and pattern recognition, pages 2285–2294, 2018. 1, 2


[53] Ziyu Wei, Xi Yang, Nannan Wang, and Xinbo Gao. Flexible body partition-based adversarial learning for visible infrared person re-identification. *IEEE Transactions on Neural Networks and Learning Systems*, 2021. 2, 3, 4


[67] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsku Choe, and Youngjoo 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. 1, 2, 3, 4, 7, 8


[74] Liming Zhao, Xi Li, Yueting Zhuang, and Jingdong Wang. Deeply-learned part-aligned representations for person re-


