Multimodal Data Augmentation for Visual-Infrared Person ReID with Corrupted Data

Arthur Josi, Mahdi Alehdaghi, Rafael M. O. Cruz, and Eric Granger
{arthur.josi.1, mahdi.alehdaghi.1}@ens.etsmtl.ca, {rafael.menelau-cruz, eric.granger}@etsmtl.ca

Laboratoire d'imagerie, de vision et d'intelligence artificielle (LIVIA)
Dept. of Systems Engineering, ETS Montreal, Canada

Abstract

The re-identification (ReID) of individuals over a complex network of cameras is a challenging task, especially under real-world surveillance conditions. Several deep learning models have been proposed for visible-infrared (V-I) person ReID to recognize individuals from images captured using RGB and IR cameras. However, performance may decline considerably if RGB and IR images captured at test time are corrupted (e.g., noise, blur, and weather conditions). Although various data augmentation (DA) methods have been explored to improve the generalization capacity, these are not adapted for V-I person ReID. In this paper, a specialized DA strategy is proposed to address this multimodal setting. Given both the V and I modalities, this strategy allows to diminish the impact of corruption on the accuracy of deep person ReID models. Corruption may be modality-specific, and an additional modality often provides complementary information. Our multimodal DA strategy is designed specifically to encourage modality collaboration and reinforce generalization capability. For instance, punctual masking of modalities forces the model to select the informative modality. Local DA is also explored for advanced selection of features within and among modalities. The impact of training baseline fusion models for V-I person ReID using the proposed multimodal DA strategy is assessed on corrupted versions of the SYSU-MM01, RegDB, and ThermalWORLD datasets in terms of complexity and efficiency. Results indicate that using our strategy provides V-I ReID models the ability to exploit both shared and individual modality knowledge so they can outperform models trained with no or unimodal DA. GitHub code: `https://github.com/art2611/ML-MDA`.

1. Introduction

Real-world monitoring and surveillance application (e.g., individuals in airport, and vehicles in traffic) rely on challenging tasks, like object detection [55, 51], tracking [29], and re-identification (ReID) [24, 50]. The aim of person ReID is to recognize individuals over a set of distributed non-overlapping cameras. State-of-art systems for person re-identification (e.g., deep Siamese networks) typically learn an embedding through various metric learning losses, which aim at making similar image pairs (with the same identity) closer to each other and dissimilar image pairs (with different identities) more distant from each other. Despite the recent advances with deep learning (DL) models, person ReID remains a challenging task due to the non-rigid structure of the human body, the different viewpoints/poses with which a person can be observed, image corruption, and the variability of capture conditions (e.g., illumination, scale, contrast) [3, 31].

Visible-infrared (V-I) person ReID aims to recognize individuals of interest across a network of RGB and IR cameras. IR cameras are often employed in conjunction with RGB cameras for, e.g., night time recognition in outdoor environments. Most approaches for V-I person ReID focus on the cross-modal matching problem. This paper focuses on person ReID systems that allow for fusion of visible and infrared modalities based on a joint representation space. Although several techniques have been proposed for dynamic and attention-based fusion [23, 41], few V-I person ReID methods have been proposed for RGB-IR fusion [35]. In this setting, it is difficult to extract discriminant modality-specific features when one modality becomes corrupted, while conserving the shared modality features [2].

In real-world surveillance applications, the accuracy of person ReID models often declines when image data is corrupted by noise, occlusions, saturation, blur, weather conditions, etc. [5]. Several strategies have been developed to improve the generalization performance of person ReID models in response to corrupted image data. Using more complex DL models, trained with more data have been shown to improve the performances in object detection [32], and image classification [47] tasks. For instance, using transformer-based models may be more suitable to tackle corruption [19, 5]. However, using more complex models, like vision-transformers [12] limits real-time ReID applications. In addition, using more diverse training data can help [47], and therefore data augmentation (DA) [5] methods may improve performance, without increasing the models complexity, and while avoiding the costs of data collection and annotation [40].

In this paper, we propose a MDA strategy to improve the accuracy of V-I person ReID systems. Chen *et al*. [5] recently proposed a DA learning strategy, called the Consistent ID Loss, with Inference before BNNeck, and Local-based Augmentation (CIL). It is mainly based on local DA, and provides improvements in accuracy for unimodal (RGB) person ReID. However, the multimodal aspect has not been explored in the literature to tackle corruptions. Yet, such approach might be helpful to tackle corruption as modalities are not similarly affected by corruptions and can still benefit by DA strategies [14].

To manage corrupted image data in multimodal settings, a multimodal DA (MDA) strategy is introduced, allowing to leverage the complementary knowledge among modalities, while dynamically balancing the importance of individual modality in the final predictions. Consequently, the strategy should reduce the corruption impact. Having in mind the multimodal person ReID aspect, and regarding that person ReID datasets were only used for cross-modal ReID, protocols are provided along with a comprehensive study over three V-I person ReID datasets, SYSU-MM01 [46], RegDB [35] and (less explored) ThermalWORLD [25]. Finally, as the focus is made on corruption robustness for the multimodal setting, the corruption benchmark proposed by [5] is extended to the infrared thermal modality.

Our main contributions are summarized as follows. (1) A MDA strategy is proposed to improve the accuracy of DL models for V-I person ReID. To optimize the collaboration among modalities, discriminant joint feature representations in the DL model, our MDA strategy relies on local occlusions and global modality masking data augmentation. (2) A comprehensive V-I multimodal experimental protocol is proposed to evaluate the impact on performance of clean and corrupted image data using the well-known SYSU-MM01, RegDB, and ThermalWORLD datasets. Corruptions from [5] are extended to the infrared domain to analyse multimodal data corruption impact. (3) An extensive set of experiments is conducted, showing that the used V-I fusion model outperforms the related state-of-art models. The limitations of unimodal models are shown by comparing a basic fusion model learned with the adapted DA to the unimodal state-of-art person ReID models.

2. Related Work

A) Multimodal person ReID. Most approaches for person ReID in the last decade [50] focus on the unimodal (RGB) [38, 27] and cross-modal [13, 1, 52] settings. Few focused on combining multimodal information, despite the potential to improve performance in the joint representation setting [2]. For example, Chen *et al*. extracted contours from the RGB modality and used a two-stream CNN architecture to combine information [4]. Bhuiyan *et al*. proposed to use pose information to gate the flow of visual information through a CNN backbone [3]. These approaches used the knowledge extracted from the main modality, which would be similarly affected by image corruption.

Some approaches sought to leverage the complementarity of RGB and depth modalities for an accurate person ReID [36, 26, 30]. However, Nguyen *et al*. [35] represents the only approach where visible and infrared modalities are integrated into a joint representation space. Infrared and visual features are concatenated from embeddings extracted independently trained CNNs, and used for pairwise matching at test time. This simple model attained an impressive performances on the RegDB dataset. However, RegDB data is captured with only one camera per modality, and RGB-IR cameras are co-located, with only a single tracklet of ten images per modality and individual. For these reasons, the RegDB dataset is less consistent with a real-world scenario. In fact, the development of person ReID models that are effective in uncontrolled real-world scenario remains an open problem [17].

B) Corruption and data augmentation strategies. Data augmentation (DA) consists in multiplying the available training dataset by punctually applying transformations on training images, like flips, rotations, and scaling [7]. This way, a model usually benefits from increased robustness to image variations, and improved generalization performance. According to Geirhos *et al*. [10], training a model on a given corruption is not often helpful over other types of degradation. Yet, [39] showed that a well-tuned DA can help the model to perform well over multiple types of image corruption, through Gaussian and Speckle noise augmentation. Hendrycks *et al*. proposed the Augmix strategy [20], where various transformations are randomly applied to an image, and then mix multiple of those augmented images. Random Erasing punctually occludes parts of the images by replacing pixels with random values [53]. Those strategies allow a large variety of augmented image, simulating eventually real-world data, and hence inducing higher generalization performance.

Focusing on person ReID, Chen *et al*. [5] proposed both a corrupted RGB dataset (adapted from [18]) and the CIL learning strategy to improve systems performance under corrupted data. Their strategy is partly based on two local DA methods – self-patch mixing and soft random eras-

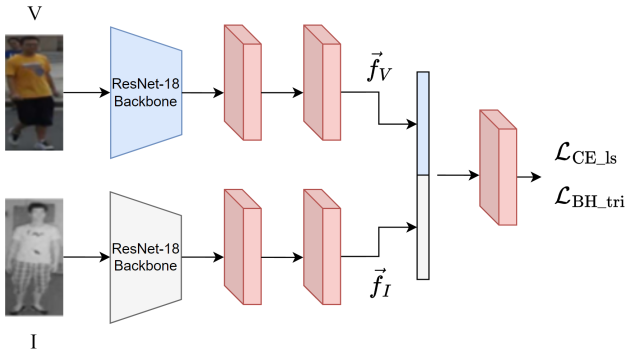

Figure 1. Training architecture considered for V-I person ReID. It learns a joint multimodal representation by concatenating features produced by independent I and V ResNet-18 CNN backbones.

ing. The former replaces some of the pixels in a patch with random values, while the latter superposes a randomly selected patch from an image at a random position on this same image. Gong *et al.* [11] show interesting improvements through local and global grayscale patch DA on RGB images. The previous strategies are limited to single modality stream models, even though the latter shows how greyscale data may reinforce the visible modality features using DA. MDA strategies have presented encouraging results for image-text emotion recognition [48] or vision-language representation learning [14]. However, to our best knowledge, our work is first to propose MDA with V-I person ReID applications.

3. Proposed Strategies

Our strategy is based on co-learning, allowing each modality stream to adapt to the other [2]. Using our MDA strategy, we expect to adapt DL models from one modality stream to another, and consequently provide better robustness to corrupted multimodal data. The low-cost multimodal architecture that we considered for V-I person ReID is based on two parallel ResNet-18 [15] backbones pretrained on ImageNet [8]. Rather than having a large single stream model, such architecture might allow us to present a competitive model both in size and efficiency. After the two backbones, each stream has an average pooling and a batch normalization layer. The final prediction is obtained by concatenating features from each embedding, right before presenting it to a fully connected layer (Fig. 1). Embeddings are concatenated during the test phase for pairwise similarity matching, from which the final ranking is obtained.

A) Multimodal patch mixing and soft random-erasing. Making a multimodal model focus on modality-specific features is challenging, as the model usually mainly focuses on shared features [2]. Augmenting data with local occlusions may help the model to emphasize modality-specific feature importance, as some features will be available only from one or the other modality.

Multimodal soft Random Erasing (MS-REA): The soft random erasing (S-REA) [5] might play this role, as it occludes parts of the RGB image punctually, potentially letting the opportunity for the hetero modality to close this occlusion gap. For S-REA, a proportion of the pixels in a given patch are given random values. To make the model close the occlusion gap in a bi-directional manner, the MS-REA is proposed (Fig. 2), applying grayscaled random pixel values on a given path of the thermal modality, as well as the random values pixel values on the RGB modality. Grayscale values respect the infrared thermal image definition as IR thermal is encoded on one channel, potentially aligning better with real-world corruptions.

Multimodal Patch Mixing (M-PATCH): Our M-PATCH DA inspired by the Self Patch (S-PATCH) DA [5]. Through M-PATCH, the idea is to extract a patch from each modality and superimpose it on the hetero-modality. The IR modality receives the RGB patch from the same individual, and vice versa. As the patches come from the same individual, the model has the option to rely on the patch features to discriminate. Three variants are explored which have different disturbance levels. From the less disturbing to the most disturbing, the first variation is extracting the patch from the Same part of the image, and applying it at the Same location on both modalities (-SS). The second extracts from the Same location but apply at Different locations (-SD), and the third extracts from Different locations and also apply at Different locations (-DD) (Fig. 2). The M-PATCH approach might gather the best of both RandomPatch [54] and S-PATCH [5] strategies. RandomPatch is strongly disturbing, and the model is forced to focus on out-of-patch features as the patch gathers information related to a different individual. S-PATCH less disturbing – it allows the model to focus on in-patch features as it contains features related to the same individual. Ours also allows in-patch feature selection by using the same individual, but provides more disturbance since the patch comes from the hetero modality. This approach may reinforce the model's shared features finding, while also pushing the model to exchange information across modalities.

B) Modality masking. A modality might be punctually unavailable or primarily uninformative. Though, the model has to know how to cancel a modality so that this one should not have a high impact in the final prediction. The modality masking approach is expected to make the model learns such behavior, by punctually replacing one or another modality with an entirely blank image. Instead of masking the multimodal representation as it has been done in [9], a representation is extracted from the masked input, so the model has to learn how to cancel its influence on the final results. This masking DA is expected to complement the

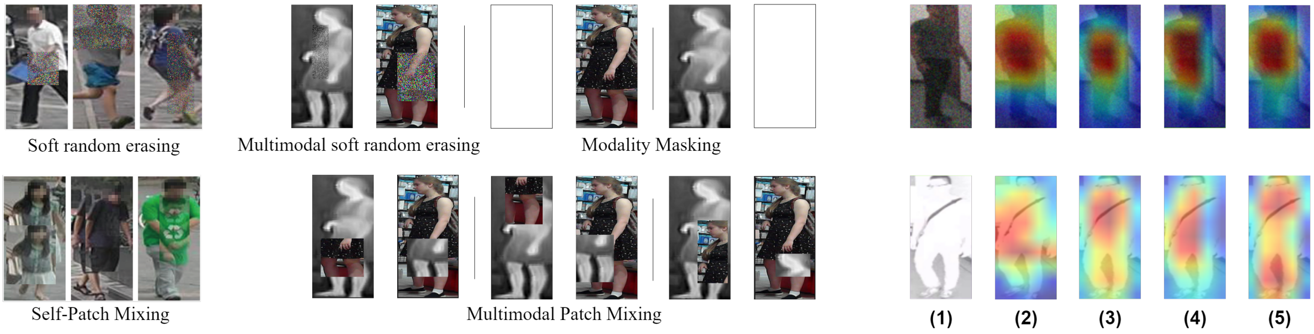

Figure 2. **Left** present data augmentation methods from [5]. **Center** are our augmentation methods extensions from those, along with the proposed Modality Masking approach. **Right** shows visualizations of activation maps from corrupted a sample and from differently trained models. **(1)** Pair input data, RGB corrupted through Gaussian noise and IR through saturation. **(2)** Augmix. **(3)** Multimodal Patch Mixing. **(4)** Modality Masking. **(5)** Multimodal soft random erasing. The discriminability increases from left to right.

previously presented DA. The M-PATCH and MS-REA approaches supposedly focus on making the model better at selecting the right features within a modality. The idea is here to balance the importance of each modality in the final embedding regarding the level of corruption of each.

4. Results and Discussion

4.1. Datasets and performance measures

Since our study is focused on V-I multimodal person ReID, we employ the widely known SYSU-MM01 (SYSU) [46] and RegDB [35] datasets, along with the lesser-known ThermalWORLD (TWORLD) [25] dataset. Details on these datasets are show in Table 1), allowing us to evaluate under diverse conditions.

Table 1. Statistics of SYSU, RegDB, and TWORLD datasets. **V**: **V**isible and **I**: **I**nfrared. Image size and number per identity is presented as: Min;Max;Avg. BRISQUE [33] metric as: avg±std.

	SYSU	RegDB	TWORLD
V-images	29,033	4,120	8,125
I-images	15,712	4,120	8,125
V-Camera	4	1	16
I-Camera	2	1	Generated
Identities	491	412	409
Paired cameras	No	Yes	Yes
V-images/id	10;144;59.1	10;10;10	1;155;19.9
I-images/id	10;144;32.0	10;10;10	1;155;19.9
Image width	26;1198;111	64;64;64	10;810;141
Image height	65;879;291	128;128;128	25;897;353
V-BRISQUE	30.50±12.26	38.84±9.86	27.79±13.28
I-BRISQUE	40.52±8.42	38.81±9.56	60.25±8.67

SYSU-MM01. [46] gather 4 RGB and 2 thermal cameras, with 491 distinct individuals, 29033 RGB, and 15712 IR thermal images. The specificity of this dataset is that its RGB and IR cameras are not co-located.

RegDB. [35] is a much smaller dataset, with one camera only per modality, co-located cameras, and a single 10 images tracklet per identity and camera. RegDB 410 identities lead to 4120 images per modality.

ThermalWORLD. [25] is only partially available, leading us to 409 distinct identities and 8125 RGB images from 16 cameras. IR images were generated synthetically. Hence, cameras can be considered as co-located. However, the thermal images are of poor quality (see BRISQUE [33] value of 60.25 in Table 1).

Corruptions. For comparison reasons, the corruptions used by Chen *et al.* [5] are the same in this study. However, the RGB corruptions were adapted to the thermal modality (detailed in supplementary material) as the thermal modality would more likely get impacted in a real scenario. The RGB data corruptions proposed by [5] are mentioned through the notation **-C**, and its extension with both modalities corrupted through the notation **-C***. Corruptions are applied independently and randomly for the RGB and the IR modalities and on both the query and the gallery images to match real-world conditions.

Performance Measures. The mean Average Precision (mAP), and the mean Inverse Penalty (mINP) are used as performance metrics, commonly used for person ReID [50].

4.2. Implementation details

Data division. SYSU-MM01 and RegDB datasets have well-established V-I cross-modal protocols [43, 44, 45, 49], but multimodal protocols remain to be built. Following SYSU-MM01 authors' cross-modal protocol, 395 identities were used for the training set, and 96 identities were used for the testing set. For RegDB, the 412 identities are kept as well into the two identical sets of 206 individuals. The SYSU-MM01 train/test ratio is kept for ThermalWORLD, leading to 325 training identities and 84 for testing. A 5-fold validation [37] is performed over the data used for training, using folds of respectively 79, 41, and 65 distinct identities for SYSU-MM01, RegDB, and ThermalWORLD.

Data augmentation (DA). The Augmix, S-PATCH, or S-REA were evaluated following the original papers settings. Our proposed multimodal extensions M-PATCH and MS-REA were used with the same appearance augmentation probability as S-PATCH and S-REA. Modality Masking is applied randomly on one or another modality, with equiprobability, and occurs with a default probability of $1/8$. For RegDB, the validation set uses the same DA as the training set. This way, better performances were observed, since they maxed out in the early epochs, or otherwise do not learn complex cues for the model.

Pre-processing. A data normalization is done at first by rescaling RBG and IR images to 144×288. Random cropping with zero padding and horizontal flips are adopted for base DA. Those parameters were proposed by [50] on RegDB and SYSU-MM01 datasets. The same normalization is kept under ThermalWORLD for consistency among protocols.

Hyperparameters. The hyperparameters values in our models were set based on the default AGW [50] baseline. The SGD is used for training optimization, combined with a Nesterov momentum of 0.9, and a weight decay of $5e-4$. Our models are trained through 100 epochs. Early stopping is applied based on validation mAP performances. The learning rate is initialized at 0.1 and follows a warming-up strategy [28]. The batch size is set to 32, with 8 distinct individuals and 4 images per individual. The paired image is selected by default for RegDB and ThermalWORLD. For the SYSU-MM01 dataset, the images from the hetero modality are randomly selected through the available ones for a given identity.

Losses. The Batch Hard triplet loss [21] $\mathcal{L}_{\text{BH\_tri}}$ and the cross-entropy with regularization via Label smoothing [42] $\mathcal{L}_{\text{CE\_ls}}$ are used as loss functions for our models. Indeed, the former is widely used in person ReID approaches [44, 6, 50], so the same margin value is fixed at 0.3, and the latter is part of the CIL implementation [5]. The total loss corresponds to the sum of both losses. The batch hard triplet loss aims at reducing the distance in the embedding space for the hardest positives while increasing the distance for the hardest negatives. The regularization with label smoothing works at reducing the gap between logits, which makes the model less confident on predictions and hence improves generalization [34].

Leave-one-out query strategy. The single-shot and the multi-shot settings [44] are widely used in cross-modal papers to form the query and gallery sets. For these settings, one or ten images from the hetero modality are selected per identity and camera to join the gallery, while the other modality forms the probe set. However, such an approach is not so realistic in a surveillance context, as the video makes the gallery number of frames per person vary much. These variations cannot be controlled as individuals are unknown in the final environment. Hence, a new strategy is developed, inspired by the leave-one-out cross-validation strategy [37], named Leave-One-Out Query (LOOQ). The LOOQ strategy treats the extreme but meaningful case in which one would have only a unique image of the person to ReID and multiple footages containing images of this same person in the gallery. Every pair of images is alternatively used as a probe set while all the other pairs join the gallery. This allows us to respect the original dataset statistics (see Table 1) by authorising the gallery images per individual to vary. Also, the mINP metric relates to the hardest test sample from the same individual. Hence, computing this metric over multiple gallery images makes it more consistent, appearing even more important in a corrupted context.

Concerning the implementation, the images are paired for both RegDB and ThermalWORLD datasets, so the paired image from the hetero modality joins the query and gallery set directly during the formation of those sets. However, SYSU-MM01 needs personal treatment since its images are not paired. Plus, the image number per modality for a given individual varies (Table 1). To solve this issue, as many pairs of images as possible are randomly selected with the constraint that one image from one modality or another must not appear in two distinct pairs. Because random image pairs are formed for SYSU-MM01, a mean of 30 trials is performed to present robust and reliable results according to the Central Limit Theorem.

4.3. Benchmarking data augmentation strategies

Table 2 shows the impact on person ReID performance of each DA strategy is investigated over the three datasets under clean and corrupted (-C*) settings. First, we compare the model learned without DA (Standard) with the model learned with Augmix, and the models learned with Augmix plus other augmentation. The other DA strategies can be S-REA, S-PATCH, or one of our proposed augmentation.

Multimodal soft random erasing. The S-REA strategy applies random values to a certain proportion of the pixels in a given patch of the RGB image. A good improvement can be seen from the Augmix to the S-REA strategy for each dataset and the clean and corrupted settings. Still, a more significant improvement happened for ThermalWORLD-C* compared to SYSU-MM01-C* and RegDB-C*, respectively, with a 11.88% improvement against 8.01% and 3.09%. While extending the DA to the multimodal setting through MS-REA, we observe a remarkable improvement for each corrupted setting, and especially that the improvement is much higher on both SYSU-MM01 and RegDB compared to ThermalWORLD. Indeed, mAP increases by 18.20% and 14.00% for SYSU-MM01-C* and RegDB-C* respectively against 3.96% for ThermalWORLD-C*. ThermalWORLD has a much weaker IR modality, so the model probably focuses much on the visible modality. Conse-

Table 2. The performance of various multimodal DA strategies using a standard model (V-I ReID model trained without DA) as baseline. Augmix DA is applied with and without other proposed DA approaches.

	SYSU		**SYSU-C***		**RegDB**		**RegDB-C***		**TWORLD**		**TWORLD-C***	
DA Strategy	mAP	mINP	mAP	mINP	mAP	mINP	mAP	mINP	mAP	mINP	mAP	mINP
Standard	96.47	73.69	25.01	1.90	99.64	98.46	21.80	2.40	87.90	49.05	29.30	3.93
Augmix [20]	95.37	68.60	35.23	2.56	99.88	99.40	40.75	9.10	87.12	46.33	42.26	5.69
+ S-REA [5]	96.21	74.36	43.24	4.06	99.90	99.51	43.84	10.25	89.24	50.10	54.14	8.92
+ MS-REA	**96.81**	**77.02**	**61.44**	**8.34**	99.86	99.35	**57.84**	**19.38**	88.95	49.92	**58.10**	**9.89**
+ S-PATCH [5]	96.40	74.89	31.39	2.14	99.90	99.53	41.83	9.39	89.12	50.53	40.73	5.63
+ MS-PATCH	94.70	69.10	33.69	2.17	99.89	99.41	40.97	9.34	**89.26**	**51.26**	41.75	5.57
+ M-PATCH-SS	96.10	73.40	35.49	2.44	99.86	99.34	43.28	10.68	88.35	50.16	44.41	5.61
+ M-PATCH-SD	95.94	72.93	35.10	2.40	99.87	99.35	42.95	10.31	88.58	51.59	43.49	5.53
+ M-PATCH-DD	94.98	68.95	33.90	2.42	99.89	99.48	41.98	9.71	88.49	51.35	43.90	5.51
+ Masking	95.61	73.49	40.92	2.90	**99.90**	**99.52**	49.27	12.10	86.01	42.76	39.91	6.16

quently, the model probably almost fully benefits from S-REA as if it were a unimodal architecture. The other datasets do not allow to benefit as much from this DA, as the model has presumably learned to focus more on IR due to the unbalanced augmentation (applied only on RGB). In contrast, the equilibrium brought by MS-REA probably allows the full exploitation of the approach and explains the impressive improvement from S-REA to MS-REA. Also, MS-REA comes first among approaches under the clean setting for SYSU-MM01 and RegDB datasets, except for ThermalWORLD. With a 95% confidence, results using MS-REA compared to the best approach are not statistically significant for RegDB, whereas it is for ThermalWORLD according to the Cochran p-values [37] of respectively 0.29 and $4.89e-5$. Thanks to MS-REA and partial occlusions, the model might have learned not to only focus on the most discriminant cues, as confirmed by the IR activation map comparison from Augmix to MS-REA (see Fig. 2). Also, this approach present important improvement over biased data augmentation, denoting a great generalization power (detailed in supplementary material).

Multimodal patch mixing. Observing the results obtained for SYSU-MM01 and ThermalWORLD, the performances globally improved from the Augmix strategy to the S-PATCH approach for the clean datasets, while those are reduced under the corrupted setting. While applying the self patch mixing on both modalities through MS-PATCH, performances are questionable, as performances remains lower or equivalent to Augmix on corrupted data, while conserving or decreasing from clean S-PATCH results. In practice, it is only while considering the modality patch exchange in our M-PATCH strategy, especially the less disturbing version M-PATCH-SS, that the best improvement is obtained on the corrupted setting, while conserving great performances on the clean one. Indeed, mAP is respectively improved by 2.15% and 2.53% over the Augmix strategy for ThermalWORLD-C* and RegDB-C*. The cameras might need to be co-located for the approach to perform, as SYSU-MM01-C* pretty much conserve similar performances as Augmix on corrupted data, and as the standard model on clean data. Spatial alignment is probably helping much the model to find correlations between the hetero modality patch and the current modality image. Still, there is a performance improvement on two datasets even if this one remains much lower than the previous MS-REA approach.

Masking. The modality masking approach presents interesting improvements under the SYSU-MM01 and RegDB datasets. Indeed, performances on corrupted datasets are increased by 5.69% mAP and 8.52% mAP over the Augmix approach, while those are pretty much matching Augmix performances on the clean datasets. The modality masking DA consists of punctually feeding a modality stream with a fully uninformative modality. Hence, those results show that the approach can make the model better able to give more importance to the discriminant modality for a given pair of images. The ability to balance modality influence on not co-located cameras through SYSU-MM01 dataset is important to highlight. Concerning the ThermalWORLD dataset, the Masking model's performances decrease for the clean and the corrupted setting compared to Augmix. Indeed, the mAP is respectively lower by 1.11% and 2.35%. Such a decrease is not surprising, as this dataset's thermal modality is very uninformative. Hence, while learning, a masked visible modality probably acts as noise by creating not discriminant V-I pairs.

Combination. As the DA approaches have distinct expected roles in the way they help the model to get more robust against corruption, combining them might allow to benefit from each of their specificities (Table 3). It is interesting to see that the real combining improvement comes

Table 3. Data augmentation combination. Each is used along with Augmix and MS-REA. C1 stands for Masking, C2 for M-PATCH-SS and C3 for Masking and M-PATCH-SS.

	SYSU		SYSU-C*		RegDB		RegDB-C*		TWORLD		TWORLD-C*	
Strategy	mAP	mINP	mAP	mINP	mAP	mINP	mAP	mINP	mAP	mINP	mAP	mINP
MS-REA	**96,81**	**77,02**	61,44	8,34	99.86	99.35	57.84	19.3	88,95	49,92	58,10	9,89
MS-REA + C1	96.77	76.01	63.01	9.59	99.90	99.45	**61.92**	**20.14**	86.34	43.24	56.10	11.04
MS-REA + C2	96.85	75.87	61.19	9.13	99.85	99.26	56.23	17.98	**89.16**	**50.68**	57.45	9.64
MS-REA + C3	96.78	75.87	**63.83**	**9.77**	**99.89**	**99.48**	61.53	20.17	86.65	43.75	**57.95**	**11.53**

from the Masking approach used with MS-REA (C1) on both SYSU-MM01-C* and RegDB-C*, with respectively 1.57% and 4.08% improvement over MS-REA used by itself. ThermalWORLD did not benefit from the masking DA, which could be expected as the Masking was already decreasing its performance when used alone. Adding M-PATCH to MS-REA (C2) or to MS-REA and Masking (C3) seems not to bring meaningful additional improvements. Indeed, MS-REA + (C3) matches the performances of MS-REA + (C1) under the clean and corrupted settings on both RegDB and SYSU-MM01. Similar observations can be done from MS-REA alone and MS-REA + M-PATCH. Hence, even if M-PATCH has shown improvements on RegDB and ThermalWORLD when used alone, those improvements are probably mainly due to the benefits of occlusions, which are already part of the MS-REA approach. Visual results observing especially IR activation maps seem to confirm this aspect (Fig. 2). Though, using MS-REA with M-PATCH appear as not being meaningful. From the previous conclusions, we propose the Masking and Local Multimodal Data Augmentation (ML-MDA) strategy, which combines both the local approach MS-REA with the modality masking DA.

4.4. Comparison with the state-of-art

Performance. As there is no true competitor in the area of V-I multimodal person ReID, the ML-MDA strategy is compared with SOTA unimodal person ReID models, with or without the CIL strategy used. According to results obtained in [5], the LightMBN [22] and TransReID [16] models are respectively the most performing unimodal models under the clean and corrupted scenarios. For fair comparison, Table 4 shows two scenarios for the multimodal test data. First, both RGB and IR are corrupted (-C*), and second, to observe how a clean IR modality can help when the RGB modality is corrupted, performance is also compared when only RGB is corrupted (-C).

Considering a clean data setting, the ML-MDA model outperforms the second-best approach by 2.32% mAP and especially by 11.22% mINP on SYSU-MM01. This significant mINP improvement shows that the multimodal setting helps considerably on the more challenging images. Indeed, the multimodal model can compensate for challenging RGB samples with the IR modality. On the RegDB dataset, our approach outperforms the others, with a statistically significant improvement. Indeed, with a 95% confidence interval, the Cochran [37] p-value between LightMBN, LightMBN + CIL and our approach is of 0.02. On ThermalWORLD, the performance of the multimodal model cannot compare with TransReID and LightMBN models, not even improving over the ResNet-18 model. Again, the poor quality IR mostly acts as a source of perturbation for the model.

When the RGB modality only is corrupted (-C) on both SYSU-MM01 and RegDB datasets, the ML-MDA model provides a considerable performance improvement over TransReID and LightMBN models. Indeed, our model reaches 87.98% mAP and 92.37% mAP for SYSU-MM01 and RegDB, respectively improving by 20.09% and 25.82% over the second-best approach. These improvements highlight the benefits of a well-trained multimodal model, relying mainly on the clean modality (I) when the other is corrupted (V). A performance gap between -C to -C* settings can be observed, as -C* is much more challenging with two corrupted modalities. The multimodal model appears as the second-best approach for both SYSU-MM01-C* and RegDB-C* datasets. Indeed, LightMBN reaches respectively 67.80% and 66.55% mAP against 63.01% and 61.92% mAP for our multimodal model. Still, the multimodal setting improves the mINP for SYSU-MM01, from 8.23% to 9.59%, and is only below RegDB by 1.39% mINP, showing that the multimodal setting can help on the hardest cases. For the -C* setting, our approach outperforms other models except for LightMBN + CIL, or on ThermalWORLD data, apparently unable to encode discriminant cues from the corrupted IR to counterbalance the well designed unimodal models. However, our architecture remains very simple, and obtaining such performance improvement on our light architecture is already promising. More complex fusion strategies, with more knowledge exchange between modality streams, and a more robust backbones like ResNet-50, may allow exceeding the performance of LightMBN.

Note that both the -C and the -C* settings might not be the most accurate for a great multimodal evaluation. Indeed, considering the IR always clean (-C) is not so accurate as weather would, for example, probably happen on both

Table 4. Performance of our multimodal model using ML-MDA compared against SOTA unimodal person ReID models, and a ResNet-18 unimodal model while using CIL or not. The two last rows show the performance of the same model when RGB is corrupted (-C), and when RGB and IR are corrupted (-C*). Note that performance on clean datasets are the same and presented in fused cells.

Model	**SYSU**		**SYSU-C**		**RegDB**		**RegDB-C**		**TWORLD**		**TWORLD-C**	
	mAP	mINP	mAP	mINP	mAP	mINP	mAP	mINP	mAP	mINP	mAP	mINP
ResNet-18	86,25	39,97	32,36	1,91	99,26	96,64	45,15	5,68	86.44	49.44	28.06	3.86
TransReID [1]	94,33	64,79	52,03	3,60	99,34	97,35	45,64	5,69	**95.86**	**77.98**	65.47	17.20
LightMBN [2]	94,45	64,06	40,90	2,13	99,90	99,41	32,40	3,25	93.02	65.94	37.34	5.60
ResNet-18 + CIL	86,64	42,78	51,64	3,83	99,65	98,41	55,76	10,98	86.95	48.07	52.85	7.97
TransReID [1] + CIL	93,20	62,02	61,38	7,20	99,69	98,57	58,74	12,89	94.79	73.82	**73.61**	**23.16**
LightMBN [2] + CIL	94,07	61,95	67,80	8,23	99,89	99,41	66,55	21,53	93.20	66.14	71.30	19.73
Ours + ML-MDA (-C)	**96.77**	**76.01**	**87.89**	**42.5**	**99.90**	**99.45**	**92.37**	**75.71**	86.34	43.24	69.20	18.47
Ours + ML-MDA (-C*)			63.01	9.59			61.92	20.14			56.10	11.04

modalities for a given co-located pair. On the other hand, considering both modalities always corrupted (-C*) hardly allows the hetero modality to help the primary modality, but is not so realistic either. Indeed, digital corruption or noise would probably not affect V-I modalities simultaneously. In fact, the real-world setting would allow [Clean RGB, Corrupted IR] pairs, and would especially be a mixture of -C and -C* settings. In practice, there should be more pairs in which one of the two modalities remains clean, so the true potential of the multimodal setting probably lies somewhere in between the -C and the -C* settings.

Table 5. Memory (number of parameters) and time (FLOPs) complexity of proposed and baseline ReID models, FLOPs computed from a single or multi-modal input.

Model	**No. Params (M)**	**FLOPs (G)**
ResNet-18	11.3	0.51
TransReID [16]	102.0	19.55
LightMBN [22]	7.6	2.09
Ours	22.5	1.54

Complexity. Multimodal person ReID with IR and RGB is more complex than regular ReID with RGB, so models are compared in terms of number of parameters and FLOPs (Table 5). The TransReID [16] model is known for being computationally expensive as its architecture is transformer based, with a total of 102M parameters and 19.55 G-FLOPs. In contrast, LightMBN [22] is based on the Os-Net architecture, which makes it very light, requires $7.6M$ parameters and $2.09G$ FLOPs. Even if our multimodal model has more parameters ($22.5M$) to adjust than LightMBN, it requires less memory compared to the SOTA unimodal person ReID models, with its $1.54G$ FLOPs. Although our model seems equivalent to LightMBN in terms of complexity, it provides a significant performance improvement. Its robustness to corrupted data makes it an excellent trade-off in the face of uncontrollable scenarios.

5. Conclusion

Real-world surveillance often requires light models that perform well on corrupted data. In this paper, image corruptions were extended to the infrared modality, and MDA strategy was proposed to improve the performance of the V-I person ReID. Experiments on the SYSU-MM01, RegDB and ThermalWORLD datasets showed the benefits of the multimodal setting over SOTA unimodal ReID models, especially when combined with the specialized MDA strategy. Indeed, our ML-MDA strategy has allowed for significant improvements in terms of robustness to corruption using the proposed modality masking and MS-REA MDA. The former learns the model to dynamically balance the importance of each modality in the final embedding. The latter works on the occlusion concept and teaches the model to better select features among modalities and not to focus only on the most discriminant features. ML-MDA improves performances, yet does not incur additional model complexity, and allows for a light ReID architecture.

Given multiple modalities, MDA allows addressing image data corruption, as these corruptions impact V and I modalities in a different ways, allowing the hetero-modality to compensate. MDA could be studied more independently from person ReID, and our methods can be applied to more general datasets (e.g., RGB-D data). Moreover, increasing the number modalities could further reduce the impact of corruption. Note that potential improvements are possible using more advanced fusion methods [41, 23]. Finally, we believe that our multimodal corrupted test set might not entirely reflect the true potential of the multimodal setting, as discussed section 4.4. To better fit real-world conditions, corruption correlations among modalities should be considered in the test set design. This would probably allow the multimodal setting to perform even better.

Acknowledgements: This research was supported by Nuvoola AI Inc., and the Natural Sciences and Engineering Research Council of Canada.

References

[1] Mahdi Alehdaghi, Arthur Josi, Rafael MO Cruz, and Eric Granger. Visible-infrared person re-identification using privileged intermediate information. *ECCV Workshops (Real-World Surveillance: Applications and Challenges)*, 2022.

[2] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.

[3] Amran Bhuiyan, Yang Liu, Parthipan Siva, Mehrsan Javan, Ismail Ben Ayed, and Eric Granger. Pose guided gated fusion for person re-identification. In *Winter Conference on Applications of Computer Vision*, pages 2675–2684, 2020.

[4] Jiaxing Chen, Qize Yang, Jingke Meng, Wei-Shi Zheng, and Jian-Huang Lai. Contour-guided person re-identification. In *Chinese Conference on Pattern Recognition and Computer Vision*, pages 296–307. Springer, 2019.

[5] Minghui Chen, Zhiqiang Wang, and Feng Zheng. Benchmarks for corruption invariant person re-identification. *arXiv preprint arXiv:2111.00880*, 2021.

[6] Seokeon Choi, Sumin Lee, Youngeun Kim, Taekyung Kim, and Changick Kim. Hi-cmd: hierarchical cross-modality disentanglement for visible-infrared person re-identification. In *Conference on Computer Vision and Pattern Recognition*, pages 10257–10266, 2020.

[7] Dan Ciregan, Ueli Meier, and Jürgen Schmidhuber. Multi-column deep neural networks for image classification. In *Conference on computer vision and pattern recognition*, pages 3642–3649. IEEE, 2012.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[9] Valentin Gabeur, Arsha Nagrani, Chen Sun, Karteek Alahari, and Cordelia Schmid. Masking modalities for cross-modal video retrieval. In *Winter Conference on Applications of Computer Vision*, pages 1766–1775, 2022.

[10] Robert Geirhos, Carlos RM Temme, Jonas Rauber, Heiko H Schütt, Matthias Bethge, and Felix A Wichmann. Generalisation in humans and deep neural networks. *Advances in neural information processing systems*, 31, 2018.

[11] Yunpeng Gong, Zhiyong Zeng, Liwen Chen, Yifan Luo, Bin Weng, and Feng Ye. A person re-identification data augmentation method with adversarial defense effect. *arXiv preprint arXiv:2101.08783*, 2021.

[12] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. A survey on visual transformer. *arXiv preprint arXiv:2012.12556*, 2(4), 2020.

[13] Xin Hao, Sanyuan Zhao, Mang Ye, and Jianbing Shen. Cross-modality person re-identification via modality confusion and center aggregation. In *International Conference on Computer Vision*, pages 16403–16412, 2021.

[14] Xiaoshuai Hao, Yi Zhu, Srikar Appalaraju, Aston Zhang, Wanqian Zhang, Bo Li, and Mu Li. Mixgen: A new multi-modal data augmentation. *arXiv preprint arXiv:2206.08358*, 2022.

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

[16] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. *arXiv preprint arXiv:2102.04378*, 2021.

[17] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *International Conference on Computer Vision*, pages 8340–8349, 2021.

[18] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

[19] Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic, Rishabh Krishnan, and Dawn Song. Pretrained transformers improve out-of-distribution robustness. *arXiv preprint arXiv:2004.06100*, 2020.

[20] Dan Hendrycks, Norman Mu, Ekin D Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. *arXiv preprint arXiv:1912.02781*, 2019.

[21] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

[22] Fabian Herzog, Xunbo Ji, Torben Teepe, Stefan Hörmann, Johannes Gilg, and Gerhard Rigoll. Lightweight multi-branch network for person re-identification. In *Conference on Image Processing*, pages 1129–1133. IEEE, 2021.

[23] Aya Abdelsalam Ismail, Mahmudul Hasan, and Faisal Ishtiaq. Improving multimodal accuracy through modality pre-training and attention. *arXiv preprint arXiv:2011.06102*, 2020.

[24] Sultan Daud Khan and Habib Ullah. A survey of advances in vision-based vehicle re-identification. *Computer Vision and Image Understanding*, 182:50–63, 2019.

[25] Vladimir V Kniaz, Vladimir A Knyaz, Jiri Hladuvka, Walter G Kropatsch, and Vladimir Mizginov. Thermalgan: Multimodal color-to-thermal image translation for person re-identification in multispectral dataset. In *European Conference on Computer Vision Workshops*, pages 0–0, 2018.

[26] Aske R Lejbolle, Benjamin Krogh, Kamal Nasrollahi, and Thomas B Moeslund. Attention in multimodal neural networks for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 179–187, 2018.

[27] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *Conference on computer vision and pattern recognition workshops*, pages 0–0, 2019.

[28] Hao Luo, Wei Jiang, Youzhi Gu, Fuxu Liu, Xingyu Liao, Shenqi Lai, and Jianyang Gu. A strong baseline and batch normalization neck for deep person re-identification. *IEEE Transactions on Multimedia*, 22(10):2597–2609, 2019.

[29] Wenhan Luo, Junliang Xing, Anton Milan, Xiaoqin Zhang, Wei Liu, and Tae-Kyun Kim. Multiple object tracking: A literature review. *Artificial Intelligence*, 293:103448, 2021.

[30] Massimo Martini, Marina Paolanti, and Emanuele Frontoni. Open-world person re-identification with rgbd camera in top-view configuration for retail applications. *IEEE Access*, 8:67756–67765, 2020.

[31] D. Mekhazni, A. Bhuiyan, G. Ekladious, and E. Granger. Unsupervised domain adaptation in the dissimilarity space for person ReID. In *European Conference on Computer Vision*, pages 159–174, 2020.

[32] Claudio Michaelis, Benjamin Mitzkus, Robert Geirhos, Evgenia Rusak, Oliver Bringmann, Alexander S Ecker, Matthias Bethge, and Wieland Brendel. Benchmarking robustness in object detection: Autonomous driving when winter is coming. *arXiv preprint arXiv:1907.07484*, 2019.

[33] Anish Mittal, Anush K Moorthy, and Alan C Bovik. Blind/referenceless image spatial quality evaluator. In *2011 conference record of the forty fifth asilomar conference on signals, systems and computers (ASILOMAR)*, pages 723–727. IEEE, 2011.

[34] Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? *Advances in neural information processing systems*, 32, 2019.

[35] Dat Tien Nguyen, Hyung Gil Hong, Ki Wan Kim, and Kang Ryoung Park. Person recognition system based on a combination of body images from visible light and thermal cameras. *Sensors*, 17(3):605, 2017.

[36] Marina Paolanti, Luca Romeo, Daniele Liciotti, Rocco Pietrini, Annalisa Cenci, Emanuele Frontoni, and Primo Zingaretti. Person re-identification with rgb-d camera in top-view configuration through multiple nearest neighbor classifiers and neighborhood component features selection. *Sensors*, 18(10):3471, 2018.

[37] Sebastian Raschka. Model evaluation, model selection, and algorithm selection in machine learning. *arXiv preprint arXiv:1811.12808*, 2018.

[38] Ergys Ristani and Carlo Tomasi. Features for multi-target multi-camera tracking and re-identification. In *conference on computer vision and pattern recognition*, pages 6036–6046, 2018.

[39] Evgenia Rusak, Lukas Schott, Roland S Zimmermann, Julian Bitterwolf, Oliver Bringmann, Matthias Bethge, and Wieland Brendel. A simple way to make neural networks robust against diverse image corruptions. In *European Conference on Computer Vision*, pages 53–69. Springer, 2020.

[40] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019.

[41] Lang Su, Chuqing Hu, Guofa Li, and Dongpu Cao. Msaf: Multimodal split attention fusion. *arXiv preprint arXiv:2012.07175*, 2020.

[42] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[43] Guan'an Wang, Tianzhu Zhang, Jian Cheng, Si Liu, Yang Yang, and Zengguang Hou. Rgb-infrared cross-modality person re-identification via joint pixel and feature alignment. In *International Conference on Computer Vision*, pages 3623–3632, 2019.

[44] Guan'an Wang, Tianzhu Zhang, Jian Cheng, Si Liu, Yang Yang, and Zengguang Hou. Rgb-infrared cross-modality person re-identification via joint pixel and feature alignment. In *International Conference on Computer Vision*, pages 3623–3632, 2019.

[45] Zhixiang Wang, Zheng Wang, Yinqiang Zheng, Yung-Yu Chuang, and Shin'ichi Satoh. Learning to reduce dual-level discrepancy for infrared-visible person re-identification. In *Conference on Computer Vision and Pattern Recognition*, pages 618–626, 2019.

[46] Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong, and Jianhuang Lai. Rgb-infrared cross-modality person re-identification. In *International conference on computer vision*, pages 5380–5389, 2017.

[47] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10687–10698, 2020.

[48] Nan Xu, Wenji Mao, Penghui Wei, and Daniel Zeng. Mda: Multimodal data augmentation framework for boosting performance on sentiment/emotion classification tasks. *IEEE Intelligent Systems*, 36(6):3–12, 2020.

[49] Mang Ye, Xiangyuan Lan, Zheng Wang, and Pong C Yuen. Bi-directional center-constrained top-ranking for visible thermal person re-identification. *IEEE Transactions on Information Forensics and Security*, 15:407–419, 2019.

[50] 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*, 2021.

[51] Syed Sahil Abbas Zaidi, Mohammad Samar Ansari, Asra Aslam, Nadia Kanwal, Mamoona Asghar, and Brian Lee. A survey of modern deep learning based object detection models. *Digital Signal Processing*, page 103514, 2022.

[52] Qiang Zhang, Changzhou Lai, Jianan Liu, Nianchang Huang, and Jungong Han. Fmcnet: Feature-level modality compensation for visible-infrared person re-identification. In *Conference on Computer Vision and Pattern Recognition*, pages 7349–7358, 2022.

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

[54] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Omni-scale feature learning for person re-identification. In *International Conference on Computer Vision*, pages 3702–3712, 2019.

[55] Zhengxia Zou, Zhenwei Shi, Yuhong Guo, and Jieping Ye. Object detection in 20 years: A survey. *arXiv preprint arXiv:1905.05055*, 2019.