IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 27 July 2024, accepted 18 August 2024, date of publication 22 August 2024, date of current version 2 September 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3447708

RESEARCH ARTICLE

Feature Compensation Network for Prototype-Based Cross-Modal Person Re-Identification

NIRMALA MURALI, (Graduate Student Member, IEEE), AND DEEPAK MISHRA, (Senior Member, IEEE)
Department of Avionics, Indian Institute of Space Science and Technology, Thiruvananthapuram, Kerala 695547, India

Corresponding author: Nirmala Murali (nirmalamurali891@gmail.com)

This work was supported by Indian Institute of Technology Palakkad Technology IHub Foundation through the Technology Development Project under the National Mission on Interdisciplinary Cyber-Physical Systems Scheme, Department Of Science & Technology (DST), Government of India, under Grant IPTIF/HRD/DF/025/SEP36.

ABSTRACT Cross-modality person re-identification is the process of matching person instances across different modalities, which poses a challenge in surveillance systems. These models encounter high intra-modality differences and a very high modality gap. In order to address the modality gap, numerous studies have employed generative models to produce image pairs as a means of augmenting the dataset. Nevertheless, the presence of artifacts in the generated images could potentially impact the model's prediction. In order to tackle these problems, we adopt a two-stage network. In the initial stage, we will extract features from cross-modality images and employ adversarial learning to produce prototype features that encompass essential features from both the RGB and IR modalities. The prototype features are utilized to construct a compensating feature set, which is then employed to train the re-identification model. The prototype features are derived from the extracted features, ensuring that only key components are utilized in the generation process. In the second phase, we employ a combination of integral probability metrics to align the identities through discrimination learning. Subsequently, we map the modalities to diminish the gap between them. At this stage, we propose to use modality-specific loss functions and modality-shared loss functions, which ensure that features relating to each modality are preserved during training. In addition, rather than identifying point-to-point differences between the feature distributions, the study focuses on the process of transporting one distribution to another, which contributes to the incorporation of perceptual learning in the model. The extensive evaluation of the proposed model demonstrates enhanced re-identification outcomes, affirming the model's ability to align modalities and augment the feature space.

INDEX TERMS Cross-modality, generative adversarial networks, integral probability metrics, person re-identification, Wasserstein distance.

I. INTRODUCTION

Person re-identification is one of the essential tasks in many computer vision applications, where a query image has to be matched with instances of a person being captured by different cameras at different places [1]. Many recent works like [2], [3], [4], and [5] have addressed person re-identification task with state-of-the-art performance. However, the re-identification system generally faces many challenges, including occlusions, viewpoint differences, low-resolution images, and illumination variations. Out of all these problems, we address the illumination variance problem, where an image of a person taken during daytime varies largely from the image taken at night. The illumination variation is a significant problem as surveillance systems must analyse data throughout the day. Today, most cameras automatically switch from RGB to IR mode at night. Even though the image set consists of the same people, the features are very different in RGB and IR domains due to the high modality gap and high intra-identity distance. Apart from

The associate editor coordinating the review of this manuscript and approving it for publication was Jiachen Yang.

this, the model also has to learn to discriminate identities within each modality. So, every cross-modal person re-identification system should handle two levels of mapping, namely, Identity Mapping and Modality Mapping.

Recent works like [6], [7], and [8] have handled cross-modal re-identification by extracting multi-granularity features. These features can be a single global feature for the entire image or some form of local feature, such as part features. Some models even estimate the pose of the person present in the image and then perform re-identification with the extracted features. Whereas few other works like [9], [10], and [11] generate paired images from the available unpaired RGB-IR images and augment these generated images to improve the feature space. Although data augmentation might be helpful as more training samples improve the model performance, the noise in the generated images might hinder the model performance. On the other hand, generating a prototype for each instance from the extracted features will enhance the feature even more and help complement the modality features. This acts as a feature compensation where the RGB and IR features are combined, and adversarial learning is performed to generate a feature prototype that compensates for the unavailability of exact RGB-IR image pairs. Recent works like [12] and [13] propose feature-level augmentation instead of image-level augmentation and then fuse these generated features. Few models that rely on metric-based learning [7], [8] are also found to be successful, where integral probability metrics are used along with the benchmark loss functions like cross-entropy loss and triplet loss. We use these works as a baseline and propose a Feature Compensation Network (FCNet) that extracts multi-granularity features and performs two-stage learning to enhance the re-identification performance. Most of the current literature focuses on compensating the images rather than the features. Therefore, we propose a model that can generate compensatory features that can be used to address the lack of precise RGB-IR pairs in the training set.

The proposed model performs re-identification in two stages; where in the first stage, a prototype feature is generated to compensate for each modality. This is done using a generative network PGAN that uses adversarial learning to combine RGB and IR features and then generate prototype features out of them. The next step is mapping the identities within each modality by finding the work done to transport one distribution to another. Later, we map the modalities in the feature space to reduce the modality gap by reducing the maximum mean discrepancy between the modality features. To reduce the high feature discrepancy between modalities, we propose a Cross-modality Batch Normalization module that can normalize the feature ranges and adapt distributions. The primary contributions of the proposed work are,

- Proposed a Prototype Feature Generation Module (PGAN) based on conditional adversarial learning, using RGB-IR features as conditional inputs to generate prototype features from unpaired images.
- Proposed a Feature Compensation Network (FCNet) with a two-step alignment process, utilizing coarse-grained and fine-grained feature mapping to learn perceptual similarity between feature distributions.
- Introduced a Cross-modality Batch Normalization (CBN) that can help in bridging the gap between RGB and IR modalities.
- Conducted comprehensive evaluations and ablation studies on SYSU-MM01 and RegDB datasets to validate the proposed method.

The paper is structured as follows: Section II discusses the related work and current literature similar to our work; Section III introduces the proposed architecture and explains the methodology in detail; Section IV analyses the results obtained from the proposed method and compares it with the current state-of-the-art models; Section V discusses the future work and areas of improvement in this project.

II. LITERATURE REVIEW

The papers [1] and [14] discuss the past and current literature. Visible Thermal person re-identification has been addressed by various methods in the past, which is discussed in [15]. We categorize the existing cross-modal re-identification methods into three types, namely, metric-based methods, feature extraction-based methods and generation-based methods.

A. METRIC-BASED LEARNING

Metric-based learning refers to learning the input images and features and then using various distance metrics to reduce the discrepancy between the ground truth distribution and feature distribution. Many metric based methods have been proposed to learn the person re-identification task in unsupervised manner [16], [17], and [18]. In the paper [19], the authors use a weighted mixup to mix the modality images linearly and leverage this as a middle modality. A center-guided metric learning is proposed to reduce the inter-modality gap. MMD re-identification model [20] uses maximum mean discrepancy as the distance metric to align the discrepancy between modalities. The paper modifies the MMD distance by adding a margin-based constraint to handle overfitting. Reference [6] proposes an adversarial learning-based model that generates modality invariant features. This improves the consistency of the features even if the modality is changed by optimizing the adversarial loss. The paper [7] proposes a data augmentation that generates HueGray images to diversify the sample space. To handle the intra-instance difference, the model uses an intra-identity representation diversification-based loss function. Reference [8] uses an informative weighted gray transform to generate gray samples in order to improve the feature space. They also use an attention module to reduce the modality gap.

Reference [21] proposes a distance metric based on the Wasserstein distance to align features based on the work done instead of point-to-point difference. To build a robust retrieval system, the authors propose a model [22] that aggregates the points and finds the class centroid representation vector used to find the triplet loss. This centroid triplet loss enhances the triplet loss function while also reducing the search space. The paper [23] proposes a triplet loss based on Wasserstein distance and also uses spatial attention to deal with the misalignment issue by finding an optimal plan to distinguish between misleading parts.

In this work, instead of relying on point-to-point distance metrics, we propose using a combination of loss functions that measure the work done to transport one distribution to another. This approach encourages the model to learn perceptual similarities between distributions.

B. FEATURE EXTRACTION METHODS

Feature extraction methods involve extracting different types of features, such as global, local, structural, and part-based features. Many models use the multi-granularity features and fuse all together for better performance. Models like [24] and [25] use a twin network that shares weights between the modalities. In [24], the authors propose a part of feature-based learning where the extracted features are divided into local feature parts and graph attention is used to align modality features. Reference [26] handles feature gaps in two levels: instance level alignment and modality level alignment. Two encoder-decoder architectures are used to learn RGB and IR features separately. Later, a prototype for each instance is built using the counterpart modality features. Reference [27] addresses the occlusion problem along with the cross-modality re-identification by using an attention mechanism and a shortest path-based algorithm that can align local and global features. The paper [19] extracts both pixel level(image) and feature level details from the images and uses both to enhance the feature space. By doing this, the modality discrepancy is handled well. Reference [28] proposes a parameter-sharing network in order to learn better discriminative features. They propose a center alignment loss that can assign features to the respective centers by a clustering process.

Unlike the above methods, we propose to use global and part features along with the generated prototype features. The generated prototype features are combined with the modality-specific features to create compensated features. Then, a two-stage feature mapping is proposed where first the identity features are mapped, and then the modality gap is mapped.

C. GENERATION-BASED METHODS

Generation-based models [29] use adversarial learning to extract or generate instance-level or feature-level information from the input. Many models that generate a middle modality have been proposed, and a new intermediate modality is generated using RGB and IR modality. One of the initial works [10] generates RGB-IR paired images to help with the insufficient images in the dataset. By doing this, instance level modality variation. Reference [30] proposed a three-mode model that generates a new modality named x-modality from the available RGB images. The x-modality is the middle modality that can bridge the modality gap between the images. Reference [12] proposes a modality-unifying network that unifies RGB and IR modalities and creates a new middle modality. A cross-modality learner is used to generate auxiliary features that have information from both modalities. Few models generate paired images for data augmentation. Reference [31] proposes a generative model where IR images are used to train the Teacher Re-ID module. The student models, along with three teacher-student loss functions, are trained to perform cross-modal re-identification. Reference [11] proposes a model that generates instance-level features as the image-level feature difference is very high. Therefore, the generated modality unifies the RGB-IR images and reduces the modality gap. Reference [9] proposes a model that generates paired RGB-IR images to compensate for the unavailability of exact RGB-IR pairs. Part-based features are extracted and used to align modality.

Reference [32] performs grayscale normalization to reduce the luminance gap. The model also extracts multi-granularity features where the head-shoulder part of the image is given more attention. The model uses both coarse-grained and fine-grained features to perform re-identification. Reference [33] proposed a non-linear generator that generates a middle modality using the features from both RGB and IR modalities. Then, the middle modality features are processed separately in RGB and IR networks, after which a distribution consistency loss is used to train the model. One of the first models to use adversarial learning for person re-identification [34], which learns discriminative features using the proposed cmGAN. Few recent works have developed feature-level generation to handle the noise in instance-level augmentation. The paper [13] discusses that generating image-level instances can induce noise into the data. Also, the re-identification network may not use the generated images effectively.

Unlike the above methods, which generate paired images, our approach generates prototype features directly. We utilize a conditional GAN (PGAN) where, instead of using labels as conditions, we input extracted RGB-IR features, which is a new way of generating features using GAN. This allows the GAN to generate prototype features encapsulating essential information from both modalities.

III. METHODOLOGY

A. OVERALL METHODOLOGY

The overall architecture is a three-stream network incorporating modality-specific features, namely, RGB, IR, and modality-shared features, as shown in Figure 1. The RGB and IR images are passed to the feature extraction network, where the modality-specific features are extracted from the convolution layer and passed to the prototype generation module (PGAN). Prototype features are generated using a conditional Generative Adversarial Network (cGAN) where RGB-IR feature vectors serve as conditional inputs. PGAN takes in the RGB and IR feature vectors and generates a prototype of the image that holds important features from both RGB and IR modalities. Then, the features in all three branches are divided into p parts to exploit the local features, which helps in learning the spatial features. The local features are then passed to a ResNet-50 block to process the features, and then the identity loss is calculated separately for each modality. Then, the generated prototype and the modality-specific features are used to learn the discriminative features and align the identities using contrastive learning. Once the identity features are learnt, the prototype and the modality-specific features are combined to get the compensated feature vectors, which are then passed to the Cross-modality Batch Normalization layer. Later, a two-step mapping, including the Identity Mapping and the Modality Mapping, is performed by using the integral probability metrics.

B. MODALITY SPECIFIC FEATURE EXTRACTION

In the three-branch network, the first and the last branch is dedicated to extracting the modality-specific features. Most models tend to mix the features to learn the shared space features. By doing this, the model might miss out on some important modality-specific features that can help improve the discriminative ability of the model. Therefore, we propose two separate branches for modality-specific features. These branches extract RGB features $F^V \in \mathbb{R}^N$, and IR features $F^T \in \mathbb{R}^N$, respectively, using the ResNet backbone. These feature extraction networks can effectively learn and represent identity-related patterns in the feature space. Modality-specific features are extracted from the last convolution layer of the ResNet block. The extracted features are then passed on to the prototype generator in the Feature Compensation module.

C. FEATURE COMPENSATION MODULE VIA PROTOTYPE GENERATION

Instead of generating image-level compensation, this work proposes to generate a feature-level prototype $F^P \in \mathbb{R}^N$. These prototype features effectively act as a representative or "prototype" of a particular instance of the identity across modalities. This improves control over the generated features, as stated in [13]. The generated feature vector will compensate for the missing information and be a prototype for the identity that holds essential feature points of RGB and IR modalities. To generate the prototype F^P, the $F^V \in \mathbb{R}^N$ and $F^T \in \mathbb{R}^N$ are extracted from the last convolution layer. Then, the features are fed into the Prototype GAN (PGAN), where the inputs are noisy latent vectors and $\{F^V, F^T\}$ feature vector pair for each instance. The feature vectors, F^V and F^T, are used as conditional labels that are used to condition the latent vector generated by the generator.

The Generator G of the PGAN has three fully connected layers stacked. The extracted features (F^V, F^T) are fed as conditional labels into the G block. These features act as conditional labels to the GAN so that the generator can condition the generated images according to the extracted RGB and IR features.

Once the feature prototypes, F^P, are generated, the feature-level prototype discriminator differentiates between the generated prototype F^P and the extracted feature pair (F^V, F^T). The discriminator then compares the generated prototype features with the real RGB and IR feature pairs, trying to fill in as many features as possible. However, since the ground truth prototype features are unavailable, we use the RGB-IR feature mixup as the ground truth here. We mix up the RGB-IR features in a linear fashion and generate $F^M \in \mathbb{R}^N$ as the ground truth, which is done as follows,

$$F^M = \lambda(F^V) + (1 - \lambda)(F^T) \tag{1}$$

where λ is a weight allocated for a domain to indicate its importance. We set $\lambda = 0.5$ so that both the modality features are mixed up equally.

The discriminator D of the PGAN has three fully connected layers stacked. The last layer gives the classification score, indicating the generated images' realness. The G and D blocks are trained adversarially so that G can improve the quality of the generated features and D can improve the discriminative capability.

$$\begin{aligned} L_{\text{GAN}} = [\log(\text{D}(F^P|(F^V, F^T)))] \\ + [\log(1 - \text{D}(F^P|(F^V, F^T)))] \end{aligned} \tag{2}$$

where D(.) is the discriminator function of the PGAN. The discriminator conditions the prototype features F^P with respect to the input features F^V and F^T. To aid in the feature generation and to align F^V and F^T better, we use the Wasserstein distance, L_W, as one of the loss functions of the GAN. The Wasserstein distance is found between the generated feature prototype F^P, and the mixup features F^M.

$$L_W = \min \sum_i \sum_j S_{ij}.M(F^P, F^M) \tag{3}$$

where S is the transport map between source i and target point j. M is the distance/ cost between the two feature distributions F^P and F^M More details on the formulation of the Wasserstein distance are discussed in the Subsection III-F1. Therefore, the PGAN loss function is a sum of the adversarial loss and the Wasserstein loss.

$$\begin{aligned} L_{\text{PGAN}} = [\log(\text{D}(F^P|(F^V, F^T)))] \\ + [\log(1 - \text{D}(F^P|(F^V, F^T)))] + \beta L_W \end{aligned} \tag{4}$$

where β is a loss weight used to indicate the importance of the L_W loss. Here, we set β to 0.5 after various experiments.

Using this module, feature-level compensation can be done, and a generic prototype can be generated, enhancing the model's shared feature learning.

D. MULTI-GRANULARITY FEATURE EXTRACTION

The FCNet model utilizes multi-granularity features to leverage the spatial characteristics of an image. This is achieved by extracting both coarse-grained (global) and fine-grained (part/local) features. Therefore, to enable shared learning, we first divide each feature vector into p parts to

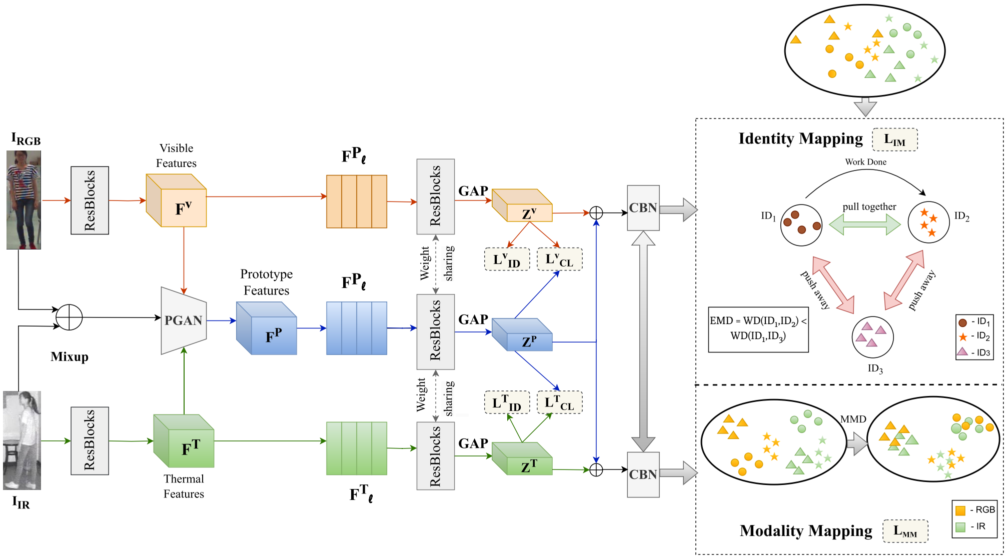


FIGURE 1. **The RGB and IR images are fed into two separate feature extraction networks, after which the features are passed to the PGAN to generate prototype features. The Mixup features F^M are given as input to the discriminator. Compensated features $\{F^V_{comp}, F^T_{comp}\}$ are created by combining the prototype features F^P_l and the modality specific features $\{F^V_l, F^T_l\}$. Cross-Batch Normalization (CBN) is applied to reduce gaps between feature distributions. A two-step mapping process is designed to map the identities and modalities.**

exploit the fine features along with the coarse/global features. This is done for all three global vectors F^V, F^T and F^P to get the local features F^V_l, F^T_l and $F^P_l \in \mathbb{R}^{N \times p}$ respectively. Then, these local features are passed onto a ResNet block, where the RGB and IR branches share weights with the prototype branch during training. To learn the discriminative features from the modality-specific features, the output of the dense layer in the ResNet block is passed onto the Global Average Pooling (GAP) to get the final activations. The GAP layer is utilized to enhance the spatial properties of the local features.

Modality-specific feature vectors are used to train the model using identity loss, L_{ID}. Therefore, two modality-specific loss functions are added to the objective function.

$$L_{ID} = -\sum_{i=1}^{N}\sum_{j=1}^{p} y_{ij}\log(C(F^C_{ij})), \text{ where, } C \in \{V, T\} \tag{5}$$

where, y is the ground truth label of the input sample, C is the trained classifier, F^C_{ij} is the feature extracted by the model C with N samples and p parts (local features) per sample.

After generating prototypes using modality-specific features, we can now proceed with contrastive learning. This involves comparing the modality-specific characteristics with the prototype features for both RGB and IR modalities. The objective is to reduce the contrast between every instance and the prototype features of the identity. During training, two forms of contrastive loss are calculated: one between the prototype features F^P_l and the RGB features F^V_l. The other contrastive loss occurs between the prototype features F^P_l and the IR features F^T_l.

$$L_{CL} = -\frac{1}{2N}\sum_{i=1}^{N}\sum_{j=1}^{p}\Big\{y_i d(F^P_{ij}, F^C_{ij})^2 + (1-y_i)[\max(0, m - d(F^P_{ij}, F^C_{ij})^2)]\Big\} \tag{6}$$

where $C \in \{V, T\}$, $d(.)$ is the distance between the feature distributions, y_i is 0/1 if they belong to the class or not, and m is the margin that controls the separation.

E. SHARED FEATURE FUSION LEARNING

Once the modality-specific features are extracted and the prototype features are generated for each instance of each identity, the model can now learn a modality-shared representation to get the compensated feature for each modality. This process integrates the prototype features, which fill in missing modality information, thereby bridging the gap between RGB and IR modalities. The prototype features F^P_l are fused with both RGB feature F^V_l and IR feature F^T_l branches to get compensated RGB features F^V_{comp} and compensated IR features F^T_{comp} respectively.

$$\begin{aligned} F^V_{comp} &= F^V_l \oplus F^P_l \\ F^T_{comp} &= F^T_l \oplus F^P_l \end{aligned} \tag{7}$$

where ⊕ refers to the concatenation operation.

1) CROSS-MODALITY BATCH NORMALIZATION

Batch Normalization is typically used within each layer of a neural network in order to standardize the activations throughout a mini-batch. However, in the context of cross-modality learning, where a network processes data from two different modalities, such as RGB images and IR images, conventional Batch Normalization may not effectively handle the differences in statistical characteristics between the modalities.

Therefore, Cross-modality Batch normalization (CBN) is developed to address this issue by adapting the normalizing process to accommodate the diverse statistical properties of different modalities. CBN, instead of normalizing each modality's activations separately inside a layer, analyzes the combined statistics of activations from both RGB and IR modalities. The model takes the mean and variance of one modality and uses these statistical measures to normalize the features of another modality. CBN can be calculated as follows,

$$\begin{aligned}\mathrm{CBN}(F^V_{\mathrm{comp}}) &= \frac{\gamma_1}{\sqrt{\sigma_T^2+\epsilon}}\cdot(F^V_{\mathrm{comp}}-\mu_T)+\beta_1\\ \mathrm{CBN}(F^T_{\mathrm{comp}}) &= \frac{\gamma_2}{\sqrt{\sigma_V^2+\epsilon}}\cdot(F^T_{\mathrm{comp}}-\mu_V)+\beta_2\end{aligned} \tag{8}$$

where, μ_V and μ_T are the means of the RGB and IR feature vectors; σ_V and σ_T are the variances of the RGB and IR feature vectors; γ_1 and γ_2 are the scaling factors (learnable parameters) of RGB and IR domains, respectively; β_1 and β_2 are the shifting factors (learnable parameters) of RGB and IR domains, respectively. CBN facilitates the alignment of the feature distributions in both the source and target domains, hence enhancing the performance of models trained on the source domain and tested on the target domain.

F. FEATURE MAPPING

The important issue that cross-modality re-identification faces is the alignment of RGB and IR modalities. While aligning the modalities, we also need to emphasise within modality identity alignment. Therefore, we propose a two-step alignment process, consisting of Identity Mapping and Modality Mapping.

1) IDENTITY MAPPING

The model should first learn the discriminative features of each identity within each modality. This can be done by training the model with the modality-specific features and, parallelly, training it with the generated prototype features. The modality-specific feature training will tell the model that the person looks like this in an RGB/ IR camera. The prototype features aid the model in learning modality invariant features and learning some discriminative representations of each identity. As already discussed in Subsection III-D, to discriminate modalities, first, the identity loss, L_{ID} is used.

However, the feature space is complex as the model has to learn representations from two different modalities. Discrimination learning is complex here due to the high intra-class variation and low inter-class differences. To deal with this, we propose to use the Wasserstein distance, which can align the identities closer if their features are similar based on the work done to measure.

Apart from aiding in identity learning and reducing the gap between intra-class feature space, the Wasserstein loss can bring perceptual learning to the model. Perceptual learning refers to the way humans perceive objects and distinguish between them. Wasserstein distance is an Optimal Transport metric that finds the amount of work done to transport some mass from a source point to a destination point. This gives the model a perception of which features are easily transferrable and similar. In cross-modal person re-identification, RGB image distributions need to be aligned with IR image distributions of the same identity. Also, the model must first align within modality features as a part of the Identity Mapping step. A transport plan S between features will be learnt by the model by optimizing Wasserstein distance.

Wasserstein distance associates the distance matrix M between two distributions as the work done to align one distribution to another. The Monge formulation is the basis of this notion and is defined as follows,

$$\text{Work Done} = \mu(x).M(x, T(x)), \tag{9}$$

where μ is the density of the source distribution, and M is the transport plan between the source and destination. Therefore, the Identity mapping loss L_{IM} between two feature vectors can be defined as follows,

$$L_{IM}(F^V_{\mathrm{comp}}, F^T_{\mathrm{comp}}) = \min\sum_{i=1}^{N}\sum_{j=1}^{p} S_{ij}.M(f^V_{ij}, f^T_{ij}). \tag{10}$$

where F^V_{comp} and F^T_{comp} are the compensated RGB and compensated IR feature vectors, respectively, and L_{IM} is the work done to transport F^V_{comp} to F^T_{comp}; $S(.)$ is the transport map between the source and destination features and M is cost/distance between the feature points.

2) MODALITY MAPPING

Now that the feature space has a clear representation with discrimination for each class, the model still cannot classify the samples correctly. This is because different modalities place the same identity features far apart. To bridge the gap between modalities, we propose to use the Maximum Mean Discrepancy loss function. The Maximum Mean Discrepancy (MMD) is a statistical metric used to quantify the difference between two distributions by mapping them into a Reproducing Kernel Hilbert Space. MMD is a two-sample test that takes the samples from two distributions and quantifies their difference. If the difference is smaller, then the samples belong to the same distribution. The objective function tries to reduce the gap between two identities from RGB and IR so that they stay close in the feature space. The modality mapping loss, L_{MM} can be defined as follows,

$$L_{MM}(F^V, F^T) = \sum_{i=1}^{p}(||E_V F_i^V - E_T F_i^T||^2) \tag{11}$$

where E_V and E_T are expectations over the RGB and IR domains, respectively. The above equation can be expanded as follows,

$$\begin{aligned} L_{MM}(F^V_{\text{comp}}, F^T_{\text{comp}}) &= \sum_{i=1}^{N}\sum_{j=1}^{p}\Big\{(E_V F^V_{ij})^2 + (E_T F^T_{ij})^2 \\ &\quad - 2(E_V F^V_{ij})(E_T F^T_{ij})\Big\} \\ &= \sum_{j=1}^{p}\Big\{E_V[k(f^V_{ij}, f^V_{ij})] + E_T[k(f^T_{ij}, f^T_{ij})] \\ &\quad - 2E_{VT}[k(f^V_{ij}, f^T_{ij})]\Big\} \end{aligned} \tag{12}$$

Here, k is the Gaussian kernel used for the MMD computation. The first two terms represent the within-modality comparison. The last term is the inter-modality alignment term. L_{MM} shows the proximity between these two distributions. By computing L_{MM} for every sample, the model learns how far the distributions are from each other and aligns the modalities to learn a modality invariant feature representation.

G. OVERALL LOSS FOR FCNet MODEL

The overall objective function consists of four loss functions: identity loss, contrastive loss, Identity Mapping loss and Modality Mapping loss. The overall loss equation for training the FCNet model is as follows,

$$\alpha_1 L^V_{ID} + \alpha_1 L^T_{ID} + \alpha_2 L^V_{CL} + \alpha_2 L^T_{CL} + \alpha_3 L_{IM} + \alpha_4 L_{MM} \tag{13}$$

where $\{\alpha_1, \alpha_2, \alpha_3, \alpha_4\}$ are the loss weights that control how much impact each loss function would create in the overall training process.

IV. RESULTS AND DISCUSSIONS

A. DATASET

The two widely used datasets, SYSU-MM01 [1] and RegDB [43], are utilized for training and testing. The SYSU-MM01 dataset has 491 RGB-IR image set identities captured by six cameras. Out of these, there are four cameras that capture images in RGB colour format and two cameras that capture images in infrared (IR) format. The collection has a grand total of 30,071 RGB images and 15,792 IR images. The images in SYSU-MM01 dataset include a high level of detail, however, there is a lack of exact RGB-IR pairs. The RegDB collection comprises 412 image sets of distinct individuals. There are ten pairs of RGB-IR image sets available for each identity.

B. EVALUATION METRICS

Two main evaluation metrics are used to compare the accuracy of the Person-ReID model. The first one is the rank-k accuracy. In order to determine the rank-k accuracy, top k matching samples are taken and compared with the query image. If the identity matches, then the Re-ID system correctly identifies the identities. Rank-k accuracy measures the proportion of properly identified images among the top k images predicted by the model.

Mean Average Precision(mAP) is also used for the evaluation of the model. A precision-recall graph is constructed to calculate the average precision. Multiple thresholds are applied to obtain various average precision levels. The mAP score is obtained by calculating the mean of all average precision values.

C. IMPLEMENTATION DETAILS

The input images to the model are unpaired images that are preprocessed to the shape 256 × 256 × 3. The feature extraction network uses ResNet-50 as the baseline model and adds four more layers to adapt to the current dataset. The proposed PGAN is trained for 100 epochs with an initial learning rate of 0.0001. Then after 30 epochs, the learning rate is gradually increased to 0.001. The FCNet model is then trained for a total of 130 epochs with an initial learning rate of 0.0001 and gradually increased to 0.005. While training, the training data is split into 0.8 and 0.2 as training and validation data. RMSProp optimizer is used to achieve convergence faster. The loss weights $\alpha = \{\alpha_1, \alpha_2, \alpha_3, \alpha_4\}$ are set to $\{0.3, 0.3, 0.5, 0.5\}$ for both SYSU-MM01 and RegDB datasets. The model is implemented in Tensorflow and trained on the NVidia Tesla V100 card for 7 hours.

The re-ranking technique is used in person re-identification models to re-order the top-k results based on the similarity score. Here, we use the re-ranking to show the improved results of the proposed model FCNet*.

D. COMPARISON WITH THE STATE-OF-THE-ART ALGORITHMS

In this section, we compare the proposed FCNet model with a few state-of-the-art approaches such as HOG [35], LOMO [15], One-Stream [15], DGD+MSR [36], MMD-ReID [20], AGMNet [32], CM-EMD [21], CM-LSP [27], DSG [19] and CMT [26]. MMD-ReID [20] and CM-EMD [21] models are mainly compared to the proposed work as these use similar metrics. GC-IFS [9], TSME [13], AGMNet [32] and MUN [12] are a few generation-based methods that are also used for evaluation. FCNet and FCNet* models refer to the proposed models without and with re-ranking, respectively.

1) COMPARISON ON SYSU-MM01 DATASET

The results of the proposed model FCNet on the SYSU-MM01 dataset are shown in Table 1. The dataset is evaluated in two settings, all-search and indoor-search. In an all-search setting, the model is able to achieve 81.47% R1 accuracy, which is higher than the current state-of-the-art [9] by 0.77% than the current state-of-the-art method. The model achieves 79.30 mAP accuracy, which is a gain of +3% than the [26]. In the indoor search setting, the model again achieves an R1 accuracy of 85.98%, which is better than the state-of-the-art accuracy. In this setting, the model achieves 87.65 mAP accuracy. Comparing the FCNet with similar methods like

TABLE 1. Comparison of FCNet model results with state-of-the-art models on SYSU-MM01 dataset. The highlighted scores indicate the best model, and the underlined scores indicate the second-best model.

Settings	All-search				Indoor-search			
Methods	R1	R10	R20	mAP	R1	R10	R20	mAP
HOG [35]	2.76	18.25	31.91	4.24	3.22	24.68	44.52	7.25
LOMO [15]	3.64	23.18	37.28	4.53	5.75	34.35	54.9	10.19
One-stream [15]	12.04	49.68	66.74	13.67	16.94	63.55	82.1	22.95
Two stream [15]	11.65	47.99	65.5	12.85	15.6	61.18	81.02	21.49
Zero-Pad [15]	14.8	54.12	71.33	15.95	20.58	68.38	85.79	26.92
DGD+MSR [37]	37.35	83.4	93.44	38.11	39.64	89.29	97.66	50.88
D-HSME [38]	50.85	73.36	81.66	47	50.15	72.40	81.07	46.16
Align GAN [39]	42.4	85	93.7	40.7	45.9	87.6	94.4	54.3
CMM-CML [40]	51.8	92.72	97.71	51.21	54.98	94.38	99.41	63.7
DDAG [24]	54.75	90.39	95.81	53.02	61.02	94.06	98.41	67.98
MACE [41]	51.64	87.25	94.44	50.11	57.35	93.02	97.47	64.79
MMD-ReID [20]	66.75	94.16	97.38	62.25	71.64	97.75	99.52	75.95
SFANet [42]	65.74	92.98	97.05	60.38	71.60	96.60	99.45	80.05
AGMNet [32]	69.63	96.27	98.82	66.11	74.68	97.51	99.14	78.30
CMCL [44]	69.97	95.26	98.27	67.42	76.48	97.92	99.68	79.94
TSME [13]	70.34	96.75	99.26	54.36	76.83	98.84	99.89	65.02
CM-EMD [21]	73.39	96.24	98.82	68.56	80.53	98.31	99.91	82.71
MUN [12]	76.24	97.84	-	73.81	79.42	-	98.09	82.06
CM-LSP [27]	76.28	94.38	97.08	76.52	82.31	98.12	99.91	85.16
PSFLNet [28]	74.0	96.5	99.0	70.51	79.5	97.5	99.24	82.1
DSG [19]	79.25	98.69	**99.84**	64.59	85.91	**99.68**	**99.97**	76.24
CMT [26]	80.23	97.91	99.53	63.13	84.87	99.41	**99.97**	74.11
GC-IFS [9]	80.70	98.28	99.43	66.59	86.57	99.46	99.89	77.36
FCNet (proposed model)	**81.47**	**98.77**	99.59	**79.30**	**86.98**	98.49	99.35	**87.65**
FCNet* (proposed model)	**86.90**	**99.81**	**99.89**	**83.24**	**90.35**	**99.02**	**99.76**	**91.25**

MMD-ReID and CM-EMD, we see that the model has a gain of almost +10% R1 accuracy. The model achieves 2% better R1 accuracy than the DSG model, which uses both pixel and feature compensation. However, the proposed model is ranked second in R10 and R20 accuracies in the indoor search setting. The FCNet* model achieves much better R1 accuracy of 86.90% in the All-search setting and 90.35% in the Indoor-search setting.

We can observe that the overall performance of the proposed model has improved over the current state-of-the-art models. This can be attributed to the two-step mapping: Identity Mapping and Modality Mapping. Also, we use the compensated features for the FCNet training, which has enriched the feature space.

2) COMPARISON ON RegDB DATASET

The re-identification results of the proposed FCNet model on the RegDB dataset are shown in Table 2. The dataset is evaluated in two settings: Visible-to-Thermal and Thermal-to-Visible. The proposed model outperforms the state-of-the-art model PSFLNet [28] and GC-IFS [9] by 2% and 3% R1 accuracy, respectively. The model achieves a better mAP score than the GC-IFS, which is the state-of-the-art generative re-identification model. In metric-based re-identification models, the proposed model outperforms the MMD-ReID [20] and CM-EMD [21] by almost 4% R1 accuracy. The FCNet* model achieves a better R1 accuracy of 98.42% in the Visible-to-Thermal search and 96.07% in the Thermal-to-Visible search.

E. VISUALIZATION OF RE-IDENTIFICATION RESULTS

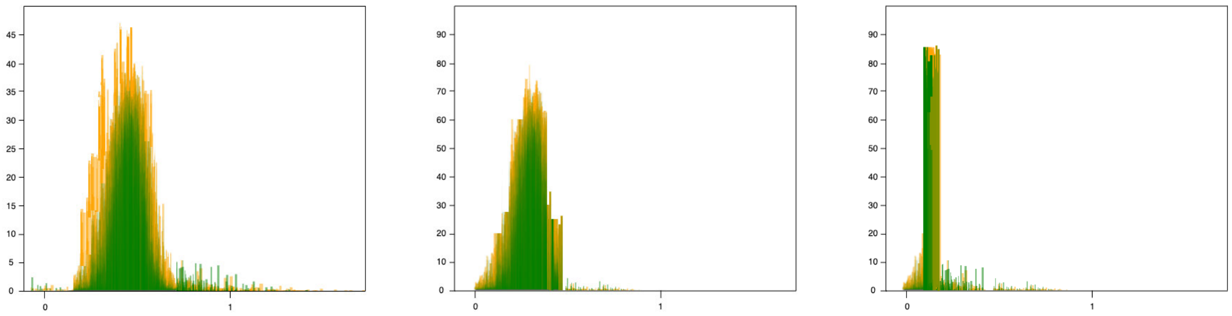
The results of the proposed FCNet model is visualized by analysing the retrieval result and the feature plots.

1) RETRIEVAL RESULTS

We plot the query and inference image list to visualise the re-identification results. Five images are randomly sampled for each identity and a query image is passed, the model returns a similarity score which is then processed to find the final output class. This is shown in Figure 3, in which the green boxes indicate the correct re-identification and the red boxes indicate that the model has identified that particular image as a different class. This is done in two settings, Visible-to-Thermal and Thermal-to-Visible. In the Visible-to-Thermal setting, the RGB image is presented as the query and inferences are made from the IR image set and vice-versa for the Thermal-to-Visible setting. It can be observed from Figure 3 that the model performs well, and very few errors occur when it is queried with an RGB image. Whereas, when the IR image is queried, there are a few more wrong re-identifications since it is a difficult setting. In the last row, we can observe that even a few back poses are correctly re-identified. This could be due to the compensated features used in the training. However, in certain low-resolution images and places with slight occlusion, the model suffers from wrong results.

2) HISTOGRAM ANALYSIS

To visualize how the feature space improves, we plot the data histogram epochwise in Figure 2. The orange distribution indicates the RGB features and the green distribution indicates the IR features. In Epoch-1, we can see that both the distributions are not aligned, and the data range is high. At the end of Epoch-30, it can be seen that the distributions started to align. At the end of Epoch-60, the feature distribution range has been reduced, and both distributions are aligned well.

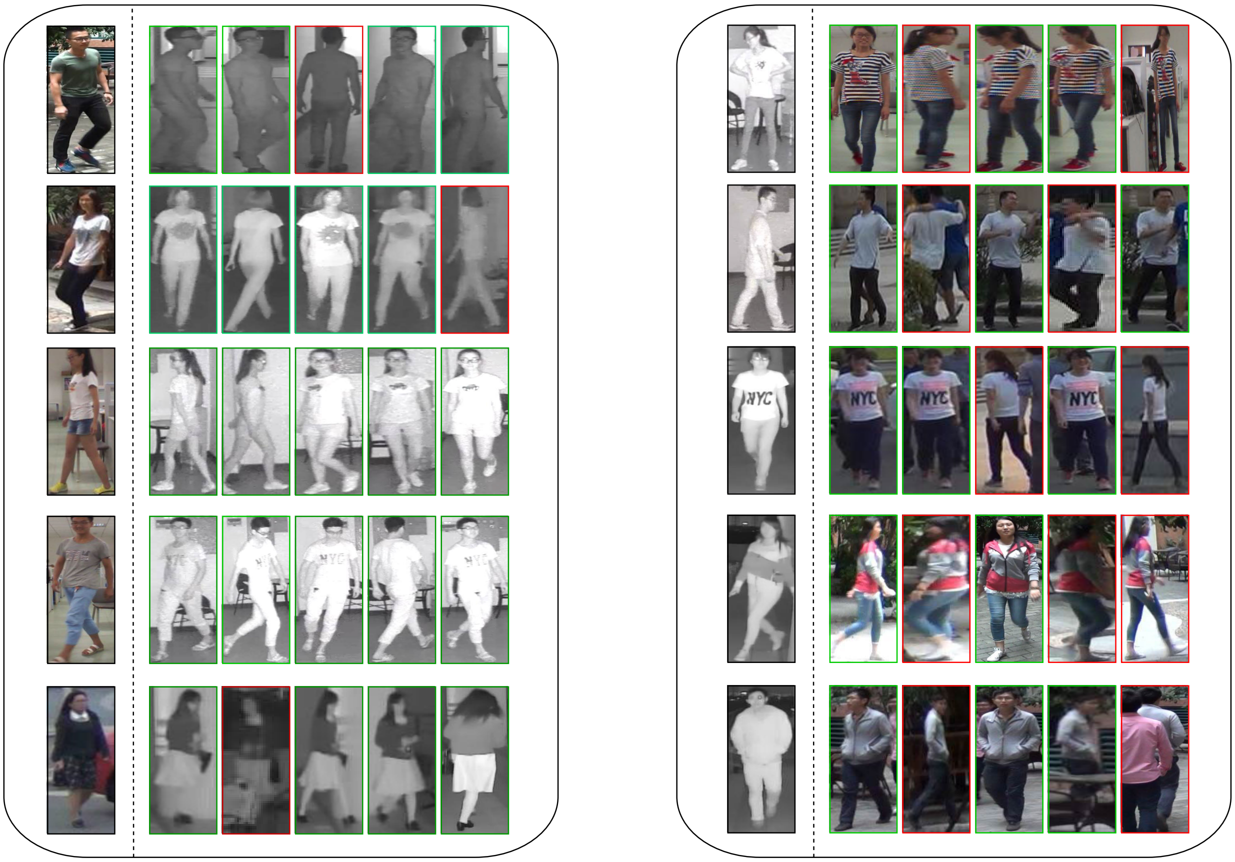


(a) Epoch-1

(b) Epoch-30

(c) Epoch-60

FIGURE 2. Visualization of the histogram alignments between RGB and IR modality through epochs in the SYSU-MM01 dataset. Orange represents RGB, and green represents the IR samples.



(a) Visible-to-Thermal

(b) Thermal-to-Visible

FIGURE 3. Image similarity check is done by randomly sampling ten identities from the gallery and trying to match images in two settings: (a) Visible-to-Thermal, (b) Thermal-to-Visible. Green boxes indicate a correct match, and red boxes indicate a wrong match.

This means that the modality gap is reducing as and when the training is done.

3) FEATURE SPACE VISUALIZATION

In order to visualize the alignment of RGB and IR feature space, we sample six identity features from the SYSU-MM01 dataset and plot the t-SNE in Figure 5. Here, Figure 5a is the result of the baseline model, and Figure 5b is the result of the FCNet model. It can be observed from the plot that the RGB and IR features are a bit apart in the baseline model. Whereas, when the features from the FCNet model are plotted, the RGB-IR features are well aligned. This proves that the proposed prototype-based FCNet model works well by aligning the modality spaces.

F. ABLATION STUDY

Within this subsection, we perform ablation studies on the proposed components of the FCNet model. To perform this study, we conduct the evaluation in two settings on

TABLE 2. Comparison of FCNet model results with state-of-the-art models on RegDB dataset. The highlighted scores indicate the best model and the underlined scores indicate the second-best model.

Settings	Visible-to-Thermal				Thermal-to-Visible			
Methods	R1	R10	R20	mAP	R1	R10	R20	mAP
HOG [35]	13.49	33.22	43.66	10.31	-	-	-	-
LOMO [15]	0.85	2.47	4.1	2.28	-	-	-	-
One-stream [15]	13.11	32.98	42.51	14.01	-	-	-	-
Two stream [15]	12.43	30.36	40.96	13.42	-	-	-	-
Zero-Pad [15]	17.75	34.21	44.35	18.90	16.63	34.68	44.25	17.82
BDTR [34]	33.47	58.42	67.52	31.83	32.92	58.46	68.43	31.96
DGD+MSR [37]	48.43	70.32	79.95	48.67	-	-	-	-
D^2RL [11]	-	-	-	-	43.4	66.1	76.3	44.1
D-HSME [38]	50.85	73.36	81.66	47	50.15	72.40	81,07	46.16
Align GAN [39]	57.9	-	-	53.6	56.3	-	-	53.4
CMM-CML [40]	59.81	80.39	88.69	60.86	-	-	-	-
DDAG [24]	69.34	86.19	91.49	63.46	68.06	85.15	90.31	61.80
AGMNet [32]	88.40	95.10	96.94	81.45	85.34	94.56	97.48	81.19
CMCL [44]	93.40	97.63	98.90	86.77	94.16	97.70	98.69	86.69
MMD-ReID [20]	95.06	99.67	99.31	88.95	93.65	97.55	98.38	87.30
CM-EMD [21]	94.37	98.93	99.42	88.32	92.77	98.50	99.66	86.85
CM-LSP [27]	94.13	-	-	88.86	93.16	-	-	87.26
DSG [19]	94.42	98.30	99.42	88.72	93.01	98.25	98.79	87.82
CMT [26]	95.17	98.82	99.51	87.3	91.97	97.92	99.07	84.46
GC-IFS [9]	94.40	**99.89**	**100.00**	92.19	92.87	99.80	**99.95**	91.00
PSFLNet [28]	95.87	98.63	99.23	91.08	92.32	97.45	98.53	88.28
FCNet (proposed model)	**97.14**	99.71	99.90	**93.03**	**95.68**	**99.86**	99.79	**92.17**
FCNet*(proposed model)	**98.42**	**99.80**	**99.89**	**94.26**	**96.07**	**99.90**	**99.87**	**93.32**

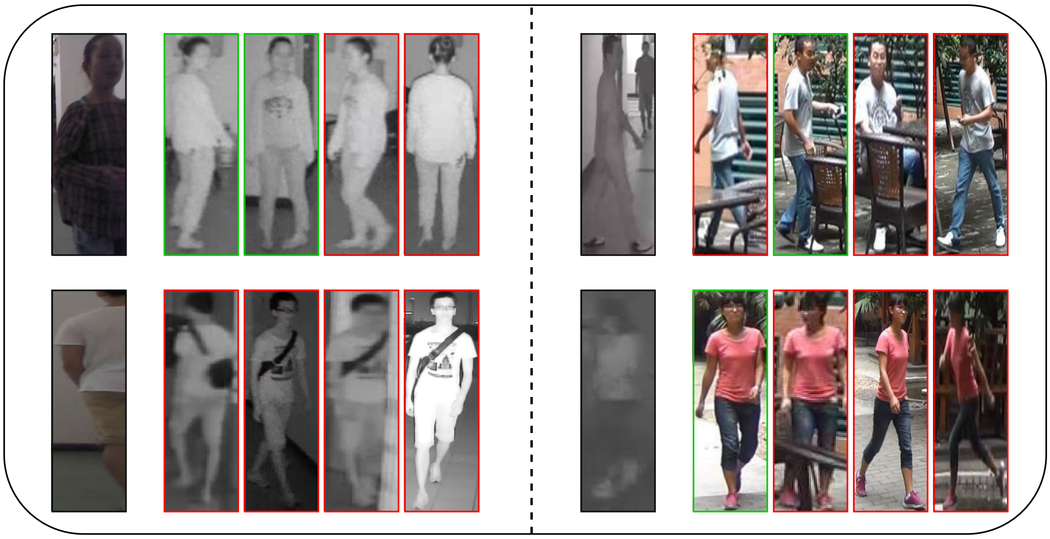


FIGURE 4. Edge case examples where the person of interest is occluded by some other object in the query image. In some cases, the query image is of very low resolution.

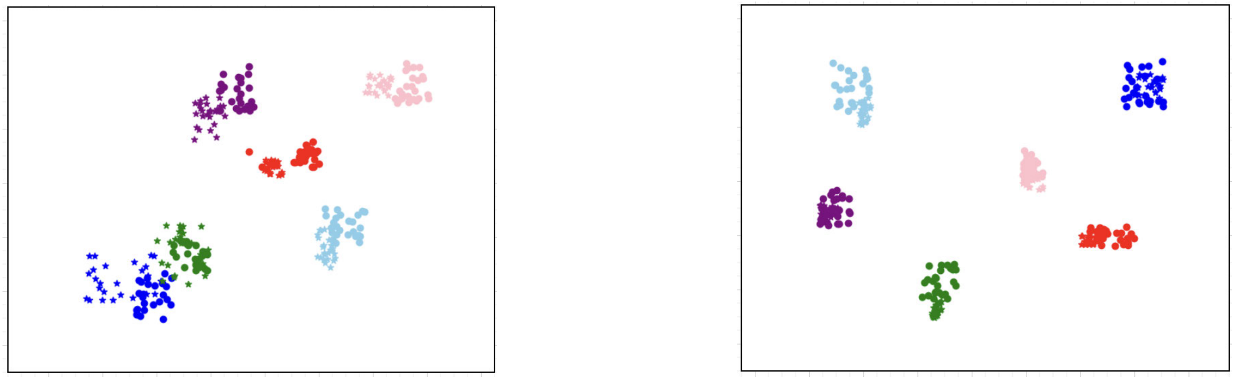
two datasets, SYSU-MM01 and RegDB dataset. The first setting checks how effective each component/module is to the model. We divide the model into four main modules: the Prototype generation module (PGAN), part features, the Feature Compensation module (FC) and Cross Batch Normalization (CBN). The second setting checks how adding loss metrics improves the model's performance.

1) IMPACT OF PGAN AND FC

In order to assess the effectiveness of the prototype features, we conduct ablation studies on the PGAN module and the FC module. As shown in Table. 3, when the PGAN module is added and when prototypes are used, the R1 and mAP are 74.19% and 74.63, respectively. However, once the PGAN module is removed, the R1 is dropped to 69.36%. In Table 4, the model has a R1 of 79.17% without the PGAN. However, once the PGAN module is added and prototypes are used, the model's R1 accuracy improves to 90.96%. This proves that Prototype Generation and Feature Compensation help improve the re-identification performance effectively. Also, the prototypes can definitely complete and enhance the feature space.

2) IMPACT OF GLOBAL AND PART FEATURES

From Table 3, when part features are not included, the R1 accuracy is 74.19%. However, once the part features are



(a) Baseline on SYSU-MM01 dataset

(b) FCNet on SYSU-MM01 dataset

FIGURE 5. t-SNE plots of extracted features from the baseline and FCNet models. Different colours indicate the six identities. The star symbol represents RGB features, and the circle symbol represents IR features.

TABLE 3. Ablation studies on SYSU-MM01 dataset.

Settings					All-search		Indoor-search	
Methods	PGAN	Part-features	FC	CBN	R1	mAP	R1	mAP
Baseline	✓		✓		74.19	74.63	77.02	80.40
		✓		✓	69.36	70.88	68.92	77.46
	✓	✓	✓		77.63	75.14	79.26	84.39
	✓	✓	✓	✓	81.47	79.30	85.98	87.65

TABLE 4. Ablation studies on RegDB dataset.

Settings					Visible-to-Thermal		Thermal-to-Visible	
Methods	PGAN	Part-features	FC	CBN	R1	mAP	R1	mAP
Baseline	✓		✓		90.96	86.90	87.10	83.66
		✓		✓	79.17	75.58	72.16	70.99
	✓	✓	✓		93.42	90.87	89.36	87.22
	✓	✓	✓	✓	97.14	93.03	95.68	92.17

used, the model can exploit the local features, and the R1 accuracy improves to 77.63%. In the Thermal-to-Visible setting, the R1 accuracy improves to 89.36%. From Table 4, we can infer that after the addition of part features, the R1 accuracy increases from 90.96% to 93.42%. However, it can be observed that only having part features will not help the model. The part features, along with the compensated features, significantly enhance the model's performance. This proves that a combination of global and part features will help in spatial understanding of the images.

3) IMPACT OF CBN

The Cross-modality Batch Normalization tries to reduce the modality gap. From Table 3, without the CBN module, the model's R1 and mAP drop down to 77.63% and 75.14, respectively. However, once the CBN is added, the model achieves a higher R1 and mAP of 81.47% and 79.30, respectively. In Table 4 as well, it can be seen that once the CBN is added, the model's R1 accuracy improves to 97.14%. This means that with the Cross-modality Batch Normalization, the distributions are aligned to the same space, therefore the re-identification performance significantly improves.

4) IMPACT OF LOSS FUNCTIONS

In this subsection, we test the influence of each loss function and combinations of loss functions on the FCNet model. The baseline model uses only L_{ID} and has an R1 accuracy of 68.36% and mAP score of 56.03, which can be observed in Table 5. Similarly, from Table 6, with the L_{ID}, the R1 accuracy is 72.95%. We propose a combination of contrastive loss L_{CL} and Identity Mapping loss L_{IM} to learn the identity discrimination. In order to verify if the identity mapping losses improve the search results, both L_{CL} and L_{IM} are added to the training. This improves the R1 and mAP to 77.49% and 75.29 in the SYSU-MM01 dataset. While testing on the RegDB dataset, it can be observed from Table 6 that the R1 and mAP increases to 93.47% and 91.63. To verify the effectiveness of the Wasserstein distance, we remove the L_{IM} and add the Modality Mapping loss L_{MM} to address the modality gap. But the performance doesn't improve as shown in the 3rd row of Table 5 and Table 6. However, once L_{IM} is included in the objective function, we can see that the R1 accuracy and mAP score increase to 81.47% and 79.30, respectively. This means that even if modality mapping is done by L_{MM}, without the L_{IM}, the high intra-identity

TABLE 5. Influence of different loss functions on SYSU-MM01 dataset.

Settings				All-search		Indoor-search	
L_{ID}	L_{CL}	L_W	L_{MMD}	R1	mAP	R1	mAP
✓				68.36	65.03	67.12	70.40
✓	✓	✓		77.49	75.29	82.53	85.69
✓	✓		✓	76.32	75.78	80.97	84.16
✓	✓	✓	✓	81.47	79.30	85.98	87.65

TABLE 6. Influence of different loss functions on RegDB dataset.

Settings				Visible-to-Thermal		Thermal-to-Visible	
L_{ID}	L_{CL}	L_W	L_{MMD}	R1	mAP	R1	mAP
✓				72.95	75.02	69.15	65.08
✓	✓	✓		93.47	91.63	92.94	88.76
✓	✓		✓	90.99	88.50	91.37	84.23
✓	✓	✓	✓	97.14	93.03	95.68	92.17

variation is not dealt with completely. Therefore, L_{CL} and L_{IM} are essential for identity mapping.

Apart from identity mapping, the important step is to reduce the modality gap by mapping the modality distributions. This is done by using the MMD loss, L_{MM}. We can see that the distributions of the RGB and IR modalities slowly align and fall within the same range from Figure 2. Also, from Table 5 and Table 6, it can be inferred that all four loss functions enhance the feature space, thereby significantly improving the cross-modal re-identification capability of the model.

5) EDGE CASES

We analyze the model's performance on some hard samples from the image set using Figure 4. All the query and inference images selected are hard samples that are either of very low resolution or occluded by some other person/object. Despite the query images being very blurry, we can see that the model still identifies two samples correctly in the first case. This could be due to the perceptual understanding brought in by the Wasserstein metric. However, in the bottom left case, since the query image is cropped, the model couldn't extract good features and hence fails to re-identify any of the images. Similarly, the query is occluded and blurred in the two cases on the right. Yet the model identifies one inference image correctly but fails to identify the rest of the samples. This means the model must be specifically trained for occlusion and viewpoint invariance.

V. CONCLUSION

A prototype-based Feature Compensation Network is proposed for addressing cross-modal person re-identification. The proposed model consists of the PGAN model that generates feature prototypes instead of generating image instances. The prototype feature generator module processes the RGB and IR features and produces a compensating feature. This compensating feature improves the feature space, followed by the utilization of a Cross-Batch Normalization module to minimize the discrepancy between the distributions of the two modalities. Subsequently, a series of loss functions are employed to align the features in two distinct phases. In the initial phase, Wasserstein loss is employed as a metric to quantify the work done between distributions and learn discriminative features. Next, the modality gap is reduced by utilizing the Modality Mapping loss in the subsequent stage. Additionally, we demonstrate that using both local and global information will improve the re-identification performance. The studies conducted using benchmark datasets provide empirical evidence that the proposed model outperforms other models in cross-modal person re-identification. The proposed FCNet obtains an R1 of 81.47%, surpassing the performance of the current state-of-the-art model. In the future, we plan to extend our research to address occlusion and various viewpoint invariances that are commonly encountered in real-world scenarios.

ACKNOWLEDGMENT

We thank the authority of IIT Palakkad Technology IHub Foundation Technology Development Project Grant No: IPTIF/HRD/DF/025/SEP36 under the National Mission on Interdisciplinary Cyber-Physical Systems scheme, DST, Government of India, for funding this research project. We also express our gratitude to the Indian Institute of Space Science and Technology for providing both technical and administrative support to this project.

REFERENCES

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

[2] J. Zuo, J. Hong, F. Zhang, C. Yu, H. Zhou, C. Gao, N. Sang, and J. Wang, "PLIP: Language-image pre-training for person representation learning," 2023, *arXiv:2305.08386*.

[3] K. Niu, L. Huang, Y. Long, Y. Huang, L. Wang, and Y. Zhang, "Comprehensive attribute prediction learning for person search by language," *IEEE Trans. Image Process.*, vol. 33, pp. 1990–2003, 2024.

[4] S. Zhang and H. Hu, "Unsupervised person re-identification using unified domanial learning," *Neural Process. Lett.*, vol. 55, no. 6, pp. 6887–6905, Dec. 2023.

[5] D. Fu, D. Chen, H. Yang, J. Bao, L. Yuan, L. Zhang, H. Li, F. Wen, and D. Chen, "Large-scale pre-training for person re-identification with noisy labels," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 01–11.

[6] J. Sun, Y. Li, H. Chen, Y. Peng, and J. Zhu, "Visible-infrared person re-identification model based on feature consistency and modal indistinguishability," *Mach. Vis. Appl.*, vol. 34, no. 1, p. 14, Jan. 2023.

[7] S. Kim, S. Gwon, and K. Seo, "Enhancing diverse intra-identity representation for visible-infrared person re-identification," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2024, pp. 2513–2522.

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

[9] J. Qi, T. Liang, W. Liu, Y. Li, and Y. Jin, "A generative-based image fusion strategy for visible-infrared person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 1, no. 1, pp. 1–16, Jun. 2023.

[10] G. Wang, T. Zhang, Y. Yang, J. Cheng, J. Chang, X. Liang, and Z. Hou, "Cross-modality paired-images generation for RGB-infrared person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 7, pp. 12144–12151.

[11] Z. Wang, Z. Wang, Y. Zheng, Y.-Y. Chuang, and S. Satoh, "Learning to reduce dual-level discrepancy for infrared-visible person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 618–626.

[12] H. Yu, X. Cheng, W. Peng, W. Liu, and G. Zhao, "Modality unifying network for visible-infrared person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 11185–11195.

[13] J. Liu, J. Wang, N. Huang, Q. Zhang, and J. Han, "Revisiting modality-specific feature compensation for visible-infrared person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 10, pp. 7226–7240, Oct. 2022.

[14] L. Zheng, Y. Yang, and A. G. Hauptmann, "Person re-identification: Past, present and future," 2016, *arXiv:1610.02984*.

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

[16] A. Wu, W.-S. Zheng, and J.-H. Lai, "Distilled camera-aware self training for semi-supervised person re-identification," *IEEE Access*, vol. 7, pp. 156752–156763, 2019.

[17] X. Jin, C. Lan, W. Zeng, and Z. Chen, "Global distance-distributions separation for unsupervised person re-identification," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 735–751.

[18] G. Chen, Y. Lu, J. Lu, and J. Zhou, "Deep credible metric learning for unsupervised domain adaptation person re-identification," in *Proc. 16th Eur. Conf. Comput. Vis. (ECCV)*, Aug. 2020, pp. 643–659.

[19] Y. Ling, Z. Zhong, Z. Luo, S. Li, and N. Sebe, "Bridge gap in pixel and feature level for cross-modality person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 6, pp. 5104–5117, Jun. 2024.

[20] C. Jambigi, R. Rawal, and A. Chakraborty, "MMD-ReID: A simple but effective solution for visible-thermal person ReID," in *Proc. Brit. Mach. Vis. Conf.*, 2021, pp. 1–20.

[21] Y. Ling, Z. Zhong, Z. Luo, F. Yang, D. Cao, Y. Lin, S. Li, and N. Sebe, "Cross-modality Earth mover's distance for visible thermal person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, vol. 37, 2023, pp. 1631–1639.

[22] M. Wieczorek, B. Rychalska, and J. Dabrowski, "On the unreasonable effectiveness of centroids in image retrieval," in *Proc. Int. Conf. Neural Inf. Process.*, 2021, pp. 212–223.

[23] Z. Zhang, Y. Xie, D. Li, W. Zhang, and Q. Tian, "Learning to align via Wasserstein for person re-identification," *IEEE Trans. Image Process.*, vol. 29, pp. 7104–7116, 2020.

[24] M. Ye, J. Shen, D. J. Crandall, L. Shao, and J. Luo, "Dynamic dual-attentive aggregation learning for visible-infrared person re-identification," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 229–247.

[25] S. Zhang, Y. Zeng, H. Hu, and S. Liu, "Noise resistible network for unsupervised domain adaptation on person re-identification," *IEEE Access*, vol. 9, pp. 60740–60752, 2021.

[26] K. Jiang, T. Zhang, X. Liu, B. Qian, Y. Zhang, and F. Wu, "Cross-modality transformer for visible-infrared person re-identification," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, Oct. 2022, pp. 480–496.

[27] X. Wang, C. Li, and X. Ma, "Cross-modal local shortest path and global enhancement for visible-thermal person re-identification," 2022, *arXiv:2206.04401*.

[28] S. Chan, F. Du, T. Tang, G. Zhang, X. Jiang, and Q. Guan, "Parameter sharing and multi-granularity feature learning for cross-modality person re-identification," *Complex Intell. Syst.*, vol. 10, no. 1, pp. 949–962, Feb. 2024.

[29] J. Liu, Z.-J. Zha, D. Chen, R. Hong, and M. Wang, "Adaptive transfer network for cross-domain person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7195–7204.

[30] D. Li, X. Wei, X. Hong, and Y. Gong, "Infrared-visible cross-modal person re-identification with an X modality," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, 2020, pp. 4610–4617.

[31] Z. Zhang, S. Jiang, C. Huang, Y. Li, and R. Y. D. Xu, "RGB-IR cross-modality person ReID based on teacher–student GAN model," *Pattern Recognit. Lett.*, vol. 150, pp. 155–161, Oct. 2021.

[32] H. Liu, D. Xia, and W. Jiang, "Towards homogeneous modality learning and multi-granularity information exploration for visible-infrared person re-identification," *IEEE J. Sel. Topics Signal Process.*, vol. 1, no. 1, pp. 1–15, Jul. 2023.

[33] Y. Zhang, Y. Yan, Y. Lu, and H. Wang, "Towards a unified middle modality learning for visible-infrared person re-identification," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 788–796.

[34] P. Dai, R. Ji, H. Wang, Q. Wu, and Y. Huang, "Cross-modality person re-identification with generative adversarial training," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 677–683.

[35] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2197–2206.

[36] Z. Feng, J. Lai, and X. Xie, "Learning modality-specific representations for visible-infrared person re-identification," *IEEE Trans. Image Process.*, vol. 29, pp. 579–590, 2020.

[37] Y. Hao, N. Wang, J. Li, and X. Gao, "HSME: Hypersphere manifold embedding for visible thermal person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, no. 1, pp. 8385–8392.

[38] G. Wang, T. Zhang, J. Cheng, S. Liu, Y. Yang, and Z. Hou, "RGB-infrared cross-modality person re-identification via joint pixel and feature alignment," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3622–3631.

[39] Y. Ling, Z. Zhong, Z. Luo, P. Rota, S. Li, and N. Sebe, "Class-aware modality mix and center-guided metric learning for visible-thermal person re-identification," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 889–897.

[40] M. Ye, X. Lan, Q. Leng, and J. Shen, "Cross-modality person re-identification via modality-aware collaborative ensemble learning," *IEEE Trans. Image Process.*, vol. 29, pp. 9387–9399, 2020.

[41] H. Liu, S. Ma, D. Xia, and S. Li, "SFANet: A spectrum-aware feature augmentation network for visible-infrared person re-identification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 4, pp. 1958–1971, Apr. 2023.

[42] P. Fang, Y. Zhang, and Z. Lan, "Beyond a strong baseline: Cross-modality contrastive learning for visible-infrared person re-identification," *SSRN Electron. J.*, p. 105, 2023.

[43] A. Wu, W.-S. Zheng, S. Gong, and J. Lai, "RGB-IR person re-identification by cross-modality similarity preservation," *Int. J. Comput. Vis.*, vol. 128, no. 6, pp. 1765–1785, Jun. 2020.

NIRMALA MURALI (Graduate Student Member, IEEE) received the B.Tech. degree in information technology and the M.E. degree in computer science from Anna University, India, in 2021. She is currently pursuing the Ph.D. degree with Indian Institute of Space Science and Technology, India. Her awards and fellowships include Ph.D. Fellowship (IIT Palakkad Technology IHub Foundation), Technology Development Project Grant under National Mission on Interdisciplinary Cyber-Physical Systems Scheme (DST, Government of India), and ACM India Anveshan Setu Fellowship (2023–2024). Her research interests include computer vision, person re-identification, and optimal transport theory.

DEEPAK MISHRA (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from Indian Institute of Technology (IIT) Kanpur, India, in 2007. From May 2007 to February 2009, he was a Postdoctoral Fellow with the University of Louisville, Louisville, KY, USA. From February 2016 to June 2017, he was a Senior IT Engineer with CMC-TCS Hyderabad, in the area of biometric research. Since 2015, he has been a Professor and the Head of Avionics at Indian Institute of Space Science and Technology (IIST), Thiruvananthapuram, India. His research interests include computer vision, biometrics, graphics, visual tracking, deep learning and its applications, and the applications of ML in the area of signal processing.