Progressive Learning in Cross-Modal Cross-scale Fusion Transformer for Visible-Infrared Video-based Person Reidentification

Hamza Mukhtar[a,b,*], Muhammad Usman Ghani Khan[a,b]

[a]Department of Computer Science, University of Engineering and Technology Lahore, G.T. Road, Lahore, 54890, Punjab, Pakistan
[b]Intelligent Criminology Lab, National Center of Artificial Intelligence, AlKhawarizmi Institute of Computer Science, University of Engineering and Technology, GT, Road, Lahore, 54890, Punjab, Pakistan

ARTICLE INFO

Keywords:
Spatio-temporal information,
Feature fusion,
Cross-attention
Multi-modality

ABSTRACT

Person reidentification (reID) is a complex problem that can be addressed by exploiting the complementary multi-modal information. We accomplish this by mapping these modalities into a shared space, which facilitates 24-hour surveillance systems. However, current methods for visible-infrared-based cross-modal person reID primarily concentrate on image-to-image matching, while the potential of image-to-video and video-to-video matching that offers a rich spatial-temporal representation, is yet to be fully investigated. Existing cross-modal reID methods rely on score fusion or feature integration techniques to merge various heterogeneous and complementary multi-modalities, unfortunately, these methods fall short of fully utilizing the complementary information offered by different modalities. To overcome the above drawbacks, this study proposes a Cross-modality Cross-scale Fusion Transformer (CMFT) for multi-scale visible-infrared complementary information interaction, yielding a more comprehensive representation for person reID. The Cross-Modality cross-scale Fusion (CCF) is a fundamental component of the CMFT, designed to capture cross-modal correlations and propagate the fused complementary and discriminative information across multiple scales. The proposed CMFT not only aligns the two modalities into a shared modality-invariant space but also captures the temporal memory to ensure motion-invariance. In order to mitigate the adverse effects of modality gaps, we propose a progressive learning scheme which first introduces the Modality-Shared Refinement Loss (MSRL) to guide the CMFT towards uncovering more reliable identity-related information from features shared across modalities. Then, a Modality Discriminative Loss (MDCL) is used to tackle the challenges of significant intra-class and minimal inter-class variation. MSRL combined with MDL enhances the discriminative power of reliable features. Importantly, our CMFT model exhibits generality and scalability, as evidenced by significant performance improvements when applied to different combinations of multimodal inputs. Experimental results validate that CMFT effectively leverages the complementary semantic information in visible and infrared inputs, outperforming the existing visible-Infrared based reID methods on the HITSZ-VCM, SYSU-MM01 and RegID datasets.

1. Introduction

Person re-identification (reID) Liu and Zhang [2019], Zeng et al. [2020], Ye et al. [2020a], Huang et al. [2022] serves a crucial role in numerous applications, including intelligent surveillance Maqsood et al. [2023], sports analytics Cioppa et al. [2022], in-store customer interaction analysis Paolanti et al. [2020]. Traditional person reID strategies strive to efficiently distinguish a target individual from a multitude of individuals present in a video stream or in images. This process of reID requires determining the correlation between two distinct images or video streams that may contain the same individual under varying conditions, such as reappearance, obstruction, different camera orientations, or distinct viewpoints. Given its central relevance to various domains, like video surveillance Maqsood et al. [2023] and customer behaviour monitoring Paolanti et al. [2020], person reID has gained significant interest. However, the mutable nature of an individual's appearance owing to diverse lighting conditions, occlusions, unrestricted postures, or changes in camera orientation, poses an ongoing and challenging issue Li et al. [2021]. The objective of reID is to associate individuals across different camera views spread over various locations Lin et al. [2022]. Although a significant proportion of present reID methods employ static images Liu and Zhang [2019], Pločo et al. [2020], Xiao et al. [2017], the trend toward video-based reID is growing Liu et al. [2019], Eom et al. [2021], Lin et al. [2022], largely attributable to the rich spatio-temporal information encapsulated in videos. Each frame in a video sequence essentially represents different manifestations of the same individual, with variations in factors like background clutter, occlusion, viewpoint, and human postures, all of which are critical to person reID Ma et al. [2021b], Ye et al. [2021b]. The majority of existing techniques for video person reID rely on supervised learning, which inherently requires access to a substantial pool of identity (ID) labelled images corresponding to each individual Lin et al. [2022].

Despite the considerable advancements made by the aforementioned methods, the majority are heavily dependent on RGB images which requires optimal lighting conditions for effective camera operation. However, this requirement can be overly restrictive, particularly in nighttime conditions where the absence of sufficient light renders the collected RGB data largely uninformative and ineffective for reID tasks. Unlike RGB images, which require sufficient light to

*Corresponding author
hamza.mukhtar@kics.edu.pk, hamza.hm.mukhtar@gmail.com (H. Mukhtar); Usman.ghani@uet.edu.pk (M.U.G. Khan)
ORCID(s):

Figure 1: The superiority of video-based cross-modal reID. Video data can offer distinguishing temporal information that is unavailable in image data, particularly when two individuals exhibit similar appearances. Specifically, under the IR camera, the person donned in the light yellowish winter coat bears a striking similarity to the individual in the black winter coat (highlighted in the green box). However, their unique arm movements provide distinctive features (indicated in the red box)).

capture details, IR images can maintain information even in low-visibility conditions. They can distinctly depict pedestrians, making them particularly useful for person reID tasks in diverse lighting conditions. As a result, RGB-IR based visible-infrared person reID Fu et al. [2021], Feng et al. [2019], Ye et al. [2020b] has emerged as a powerful strategy to facilitate reID in 24-hour surveillance. However, in most surveillance scenarios, individuals are typically represented in video databases, with each tracklet encompassing multiple frames. It is intuitive that video-based data provides a wealth of visual information that far surpasses that of a single image Aich et al. [2021]. Therefore, to fully leverage the rich information available in video data, there is a need to extend current cross-modal reID methods to handle video-based reID tasks.

In certain scenarios, distinguishing between two individuals with similar appearances based on a single image can prove to be a challenge. This difficulty is magnified in the context of the infrared modality, where even human judgment may not ensure accuracy due to the limitations in capturing colour and texture details. In contrast to static images, video sequences encapsulate both spatial and temporal information, enabling the extraction of valuable motion details that can aid in more distinctive reID. For instance, as depicted in Figure 1, two sequences of images of two individuals appear remarkably similar when viewed under an infrared camera. However, the individual wearing the light yellowish winter coat demonstrates a unique arm posture in motion that differentiates them from the individual wearing the black coat. These motion characteristics contribute more distinctive features, enabling the development of a more robust and precise reID model. As a result, it becomes crucial to prioritize videos over static images. By leveraging the rich temporal information in videos, we can extract motion-based features that offer a more comprehensive and distinctive representation of individuals. Hence, our research is focused on developing methods that can effectively capture and utilize motion-based features from video sequences for cross-modal reID.

Contrasted with single-modality reID Xiao et al. [2017], Liu et al. [2023], Lin et al. [2020], visible-infrared reID (VI-reID) Wang et al. [2019a], Ye et al. [2020a], Wang et al. [2022b] presents three primary challenges: 1) Modality Gap: The considerable gap between the two modalities creates a significant challenge when trying to harmonize the identity-related features across them. Each modality captures different aspects of an individual's appearance, with visible imagery primarily capturing colour and texture details and infrared imagery capturing thermal information. This difference in the nature of the features captured by each modality can make it challenging to align the identity-related features across them. 2) Lighting Variations: Infrared images are more susceptible to changes in lighting conditions compared to visible images. This sensitivity can lead to less discriminative features for cross-modality matching. While visible images are largely dependent on external light sources, infrared images are based on thermal radiation produced by the person's body. This can make infrared images more sensitive to changes in environmental and body temperatures, resulting in less consistent and less discriminative features. 3) Changes in Clothing: Changes in clothing based on modality can occur due to the extensive time span between the capturing of visible and infrared images. This can increase the complexity of robust feature extraction. While

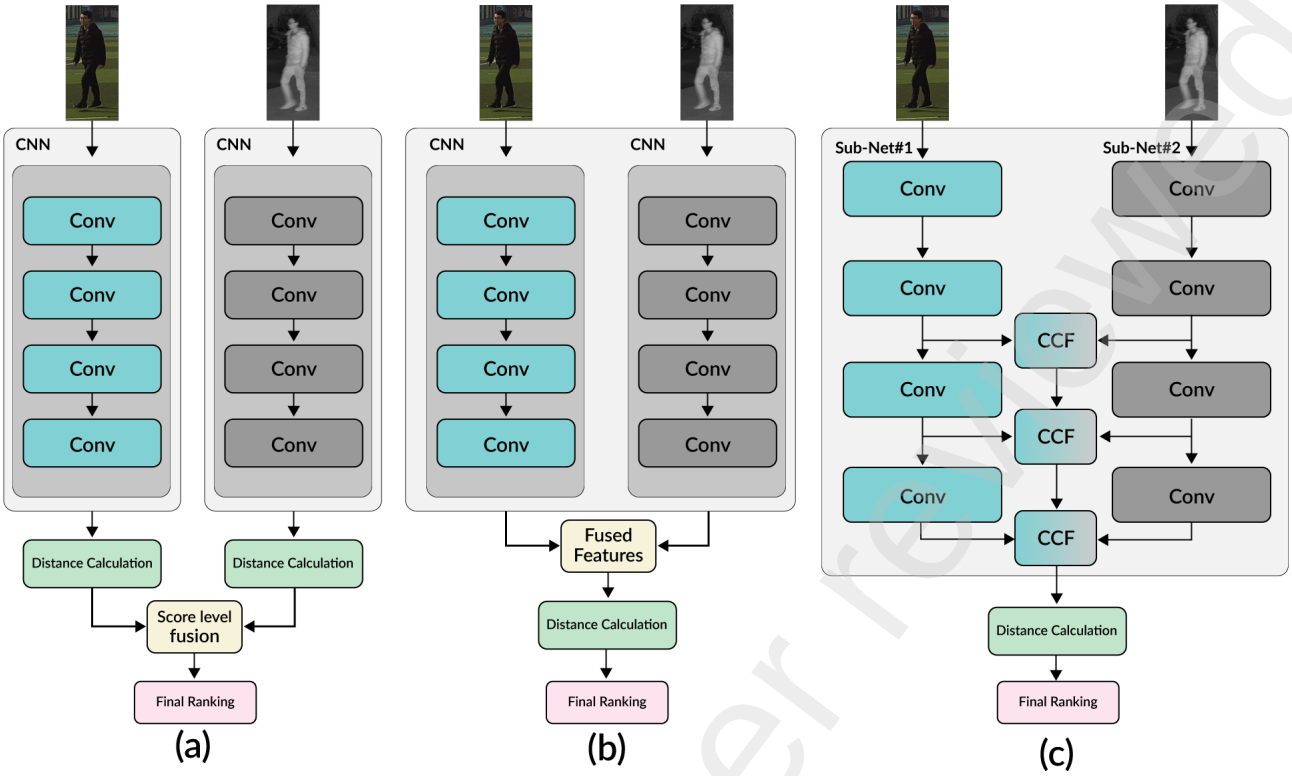

Figure 2: Visual representation of three distinct multimodal fusion methods. (a) Score fusion, which is a method of combining the output vectors to yield a singular score. (b) Single-scale feature fusion, which involves the integration of features at a single scale. (c) Cross-modality feature fusion incorporating our proposed CCF that gathers and learns multimodal information across varying feature scales between dual-stream CNN.

visible images capture the colour and pattern of clothing, infrared images only capture the heat signature. Therefore, any changes in clothing between the capturing of the visible and infrared images can introduce inconsistencies and make it more difficult to extract robust and consistent features. These challenges highlight the complexities involved in VI-reID and underscore the need for innovative and robust methods that can effectively handle these complexities.

Despite the potential advantages, existing research Xiao et al. [2017], Liu and Zhang [2019], Zhang et al. [2020], Liu et al. [2023], **?** has primarily focused on leveraging a single modality for person reID dynamics modelling and reasoning about spatiotemporal relationships. These methods do not fully utilize the complementary features offered by multiple modalities. The research focus has been largely skewed towards modalities such as RGB, optical flow, and skeleton, while other promising modalities, such as infrared, have been largely overlooked. Although some studies Vaswani et al. [2017], Li et al. [2021], Zhang et al. [2021a], Chen et al. [2022], Tang et al. [2023] have made efforts to incorporate multiple modalities, the extraction of more discriminative and effective multimodal representations for person reID continues to be a complex and unresolved issue. As depicted in Figure 2(a), the initial approach uses score fusion for merging the visible and infrared input and combining prediction scores from separate CNNs that have been trained on separate modalities as classifiers first and then uses as feature extractors. However, this basic score fusion Yoon et al. [2023] approach fails to consider the multi-modality feature interactions and correlations and does not fully utilize complementary cross-modality features. Some studies Lu et al. [2023], Jiang et al. [2022] have recognized this limitation and have fused features obtained from each single-modality branch within a network trained jointly, as shown in 2(b). Despite these efforts, the potential of multimodal fusion in person reID is yet to be fully realized and presents an area for further exploration and improvement. The fusion methods described earlier mainly focus on the integration of high-level semantic features, neglecting the delicate complexities that exist across different modalities and feature scales. These approaches miss the opportunity to exploit information from various scales, which can lead to a more richer understanding of the underlying patterns. The central goal should instead be to leverage multi-scale features in order to create richer and more comprehensive multimodal representations, capturing correlations and interconnections that might otherwise be overlooked. Moreover, an overly simplistic fusion strategy can occasionally act as a detriment to the performance of multimodal learning Lu et al. [2023]. This raises the need for a more comprehensive approach to multimodal fusion that takes into account the diverse scales of features. The ultimate goal is to develop a strategy that can

effectively manage and integrate multi-scale features from multiple modalities, leveraging the strengths of each modality to improve the reID. In response to these challenges, we propose the Cross-modality Cross-scale Fusion Transformer (CMFT), a progressive cross-modality fusion technique, to efficiently capture and utilize the complementary information present in both visible and infrared modalities. A key element of the CMFT is the Cross-modal cross-scale fusion (CCF), which is specifically engineered to facilitate exchanges between visible and infrared modalities across multiple CNN blocks. We use dual-stream CNNs to capture spatiotemporal information which are progressively merged at various scales using a joint training approach with the CMFT. This approach ensures that the model can effectively capture and integrate the unique and shared characteristics present across different scales and modalities. A visual depiction of our multimodal fusion technique is provided in 2(c).

In video-based cross-modal reID, spatial information provides critical cues about the appearance of the individual, while the temporal information encapsulates the motion patterns over time. These two types of information, when combined, offer a more comprehensive understanding of the individual's identity. This is especially true when considering cross-modal scenarios, where the individual is observed through different modalities such as visible light and infrared. By effectively capturing and integrating spatial and temporal information from each tracklet, we can obtain a more discriminative and robust representation for each individual. In this study, we first introduce a progressive learning strategy that employs Transformers Wang et al. [2019a] aimed at minimizing the disparity between the visible and infrared modalities Zhang et al. [2023]. We specifically enhance the hard triplet loss, which aids in distinguishing between different identities, enhancing the model's ability to differentiate subtle differences. We also incorporate visible images as a supplementary modality to capture modality-independent patterns, making our model robust against variations in the input data. This cross-modal learning approach allows the model to identify and utilize patterns that are consistent across different modalities. Moreover, we propose a novel Modality-Shared Refinement Loss (MSRL) which alleviates the adverse effects that arise from discrepancies between modalities and strengthens features that contain information shared across both visible and infrared modalities. By focusing towards these shared features, MSRL allows for more effective use of the complementary information present in both modalities. Alongside this, we introduce a Modality Discriminative Loss (MDL) to address the prevalent issue of large intra-class variations. This loss function enhances the distinctiveness of reliable modality-shared features, ensuring a greater similarity among features belonging to the same class compared to those from different classes. This decrease in intra-class variability ultimately improves the model's capacity for discrimination. The combined application of these strategies provides a robust approach to tackling the inherent complexities of cross-modal reID, resulting in improving the performance of this task.

Key contributions of this study are:

1. We introduce the CCF that is designed to carry out the interaction and propagation of multimodal features across various scales. This concept is specially desinded for Visible-Infrared-based person reID tasks, thereby providing a new perspective on multimodal feature extraction and fusion.
2. we arrange CCFs at distinct feature levels to construct the Cross-modality Fusion Transformer (CMFT). which demonstrates its universal applicability for multimodal person reID tasks, as it can be extended to accommodate various combinations of multimodal input. This universal applicability makes our approach highly versatile and adaptable to different multimodal reID tasks.
3. We propose a unique Modality-Shared Refinement Loss (MSRL), specifically designed to amplify the modality-common features, effectively resolving the issue of feature unreliability. By focusing on the common features, MSRL allows the CMFT to better leverage the complementary information present in both modalities and improve the reliability of the extracted features.
4. We introduce a novel Modality Discriminative Loss (MDL) to handle substantial intra-class variations that boosts the discriminative capacity of features that are invariant to the modality. By reducing intra-class variations, MDL enhances the discrimination ability of the features and improves reID accuracy.
5. By integrating our CMFT with dual-stream CNNs for processing the frame sequence, it has shown promising performance compared to existing Visible-Infrared based methodologies. This is evidenced by our comparative analysis of three commonly utilized datasets: HITSZ-VCM Lin et al. [2022], SYSU-MM01 Wu et al. [2017] and RegID Park et al. [2021].

The remaining article is organized as follows: Section 2 reviews the different types of previous reID methods such as single-modality-based reID, Multi-modality based reID, Vision-Transformer based reID and Video-based reID. Section 3 presents our proposed progressive CMFT, while training, validation and performance comparison on public benchmark datasets are carried out in Section 4, followed by Section 5, which summarizes the research and provides the future research direction.

2. Related Work

The studies relevant to our work are those focusing on multimodal person reID and different cross-modality complementary techniques. Person reID Pločo et al. [2020], Zeng et al. [2020], Wang et al. [2019a], Ye et al. [2020a], An and Liu [2022] determines the presence of particular individuals across monitoring devices with varied spatial

distributions. The concept of cross-modal reID Wang et al. [2019a, 2020] has emerged as a pivotal and challenging area within the computer vision field, generating substantial research interest. This section discusses various signal-modality and cross-modality, visible-infrared person reID methods where cross-modality methods mainly focus on the Visible-Infrared modalities. Then transformer-based reID methods are discussed to show the potential of the attention mechanism. Moreover, video-based reID are also discussed which is the main focus of our research.

2.1. Single-Modality based Person reID

Person, reID utilizing a single modality Liu and Zhang [2019], Lin et al. [2020] strives to identify matches for images of individuals captured by different visible cameras. This task is inherently complex due to substantial variations that can occur in perspective and human pose when images are captured from different camera viewpoints. To address these complexities, existing methodologies have primarily focused on two areas: representation learning Liu and Zhang [2019], Zhang et al. [2020], Pločo et al. [2020] and metric learning Xiao et al. [2017], Liu et al. [2023], Lin et al. [2020], Zheng et al. [2023]. The former focuses on deriving salient and discriminative features from the images, while the latter is concerned with devising distance metrics to evaluate the similarity between these feature representations. These two approaches have been instrumental in significantly improving the performance of benchmark datasets in the field of person reID. Despite these advancements, there remains room for further exploration and improvement, especially in the context of multi-modal person reID which introduces additional complexities such as inter-modality discrepancies. From the feature representation perspective, one work Zhang et al. [2020] introduced an end-to-end learnable architecture to learn view-agnostic identity-specific attributes. Other work Pločo et al. [2020] developed a spatial-temporal constraint to shape a distribution, thus achieving efficient feature representation. In study Liu and Zhang [2019], the authors proposed a fusion process at the feature map level using the attention mechanism, which provided a more robust feature representation. Regarding distance metric learning, the goal is to design models that guarantee a shorter distance between true matches compared to incorrect ones. The metric learning method, referred to as Margin Sample Mining Loss Xiao et al. [2017], selectively computes the final loss based on positive pairs with the largest distance and negative pairs with the shortest. Generative metric learning Liu et al. [2023] aimed to enhance open-world reID by generating adversarial examples and target variants, which mitigate the effects of disturbance attacks and ensure accurate predictions in the feature metric space. Furthermore, some studies Wang and Zhang [2020] treated the reID as a multi-class classification, categorizing images of the same individual into a singular category. Moreover, a triplet loss is also used to stimulate the proximity of representations acquired from the same identity, while prompting representations from disparate identities to occupy separate spaces within the feature domain Zeng et al. [2020]. However, the aforementioned studies predominantly tackle single-modality matching issues, which restricts their practical applicability, particularly in surveillance contexts with insufficient or no lighting. Unlike these studies, our proposed method introduces CFMT to develop a unified reID framework that effectively manages the complementary RGB and infrared information.

2.2. Visible-Infrared Modality based Person reID

Visible-infrared reID involves the matching of individuals using the input of multiple modalities, specifically RGB cameras and IR cameras. Existing methods Wang et al. [2019a], Ye et al. [2020a] largely concentrate on learning shared feature representations across modalities to counter this challenge. Some image translation-based methods Wang et al. [2019a, 2020], Ye et al. [2020a], Fu et al. [2021] aim to unify the modalities before learning shared representations. For instance, this work Wang et al. [2019a] proposes an alignment generative adversarial network (GAN) that can learn end-to-end and combines pixel-level and feature-level alignment. Another study Wang et al. [2020] generates paired images across modalities and performs alignment at both the global set level and the fine-grained instance level. An alternative approach involves designing various dual-stream architectures Ye et al. [2020a], Fu et al. [2021] to capture both modality-specific and shared features. For instance, this study Ye et al. [2020a] present a modality-aware collaborative learning pipeline with a dual-stream network that can share parameters.Fu et al. [2021] utilizes neural architecture search to identify an optimal dual-stream architecture for VI-reID. However, these methods often overlook modality-specific features, limiting the discrimination ability of the learned representations. To overcome this limitation, modality compensation methods Wang et al. [2019b], Lu et al. [2020] have proposed that aim to compensate for the lack of specific modality features. For example, one study Wang et al. [2019b] generates multi-spectral images to compensate for the absence of specific information by using GANs. These methods seek to maintain the unique characteristics of each modality while still achieving robust cross-modality in reID.

The study in Lu et al. [2020] proposed an algorithm for transferring features shared across modalities, taking advantage of both shared and modality-specific features. However, the compensation process in this method is dependent on the current mini-batch, leading to variability in the generated modality features across different mini-batches. Nevertheless, these methods did not completely reduce the cross-modal discrepancies. In response, different methods have adopted a cross-modal triplet loss where the anchor and positive/negative pairs are sampled from person images across different modalities Feng et al. [2019], Wang et al. [2022b]. For instance, in these methods, RGB images are often utilized as anchors, while infrared images are served as positive or negative samples. The idea behind this approach is to prompt the features extracted from images of the same

individual, although from different modalities, to mirror one another. The result is the construction of person representations that are not only highly discriminative but also resistant to the variations introduced by the use of different modalities. Consequently, this makes these representations more robust, enabling them to effectively resist against the inconsistencies and discrepancies typically associated with cross-modal data. Most recently, the DDAG Ye et al. [2020b] experimented with the use of a graph attention network to directly manage the relationships between RGB and infrared images. This method takes a more explicit approach to cross-modal discrepancy reduction, using a graph-based model to capture and leverage the relationships between modalities. Unlike the existing method, our approach introduces a dual-stream sequence-based feature extractor instead of 2D CNN which provides spatial-temporal features, and then these features are fed to CCF at different levels for cross-modality complementary discriminative information.

2.3. Transformer based Person reID

Transformer-based models Lai et al. [2021], Zhang et al. [2021a]based on the multi-head self-attention, have demonstrated a robust capacity for capturing global dependencies Vaswani et al. [2017]. Recent trends have shown great interest in applying the transformer architecture to a broad range of computer vision tasks, such as image classification Dosovitskiy et al. [2020], object detection Carion et al. [2020], and image segmentation **?**. However, in the context of reID, the current approaches involve deploying Transformers for a single modality. For example, Transreid He et al. [2021] utilized a pure-transformer model along with a side information embedding and a patch module, to extract discriminative features. PAT Li et al. [2021] used a transformer network for capturing the various occluded parts of the person for reID. DTF Zhu et al. [2021a] implemented the learnable vectors of part tokens to capture part features, and incorporated part alignment within the self-attention mechanism. TPM Lai et al. [2021] adopted the transformer to facilitate adaptive part divisions. Another study Zhang et al. [2021a] introduced a hierarchical aggregation Transformer to combine multi-level features. HAT Chen et al. [2021a] combines the CNN and attention mechanism for hierarchical aggregation features extracted from a CNN. Oh-former [35] combines the CNN and transformer for extracting the high-order correlation.

Furthermore, PIRT Ma et al. [2021b] has crafted a pose-guided intra-relational and inter-part relational Transformer specifically designed for capturing the long-term correlation. CMT Jiang et al. [2022] used a Transformer encoder-decoder mechanism to compensate for any absent modality-specific features and adjust the instance-level features. Cross-modality SPOT Chen et al. [2022] leveraged person body pose point data and devised a structure-aware positional Transformer to extract features that are both semantic-aware and shared across modalities. While these Transformer-based approaches have indeed enhanced, we explore the potential of progressively employing modality-shared Transformers to learn reliable features that are pertinent to the task of Visible-Infrared reID. Slightly deviating from prior approaches, our CMT extracts the features from the different stages of sequence-based feature extractor and feeds for the interaction and fusion of both modality features through the progressive CCF with the key objective of making up for the absence of modality-specific information.

Transformers have been employed for person reID in two primary ways. Some approaches integrate transformers above convolutional layers for temporal and relational reasoning Lu et al. [2023], Jiang et al. [2022], while others substitute CNNs with a transformer-based backbone to extract spatiotemporal context Bertasius et al. [2021], Arnab et al. [2021], demonstrating impressive outcomes on video-based human activity recognition. For instance, the Video Swin Transformer Liu et al. [2022] employs a local self-attention mechanism, which offers a computationally efficient and compact model-size approximation for global spatiotemporal self-attention. However, Uniformer Li et al. [2022] tried to overcome the inability of self-attention to efficiently learn low-level information. Instead, it proposed a model integrating 3D convolution with self-attention for streamlined video representation learning. Contrary to traditional self-attention-based temporal modelling methods that consider identical spatial positions across frames to capture dynamics.

In addition to the conventional methods of temporal feature learning, there are two other approaches that have been proposed in recent studies. One of these is novel inter-frame attention to measure the temporal correlation within a localized region Long et al. [2022] Another approach is the implementation of the dynamic temporal filter, which encodes spatially-aware temporal representation in the frequency domain Long et al. [2022]. This mechanism can be integrated into transformers to extend the temporal receptive field. Contrary to the prior studies, our proposed CMFT method adopts a progressive approach to incorporating and propagating multimodal data across varied feature scales within a singular path. This path is trained in coordination with unimodal network paths. In this model, the transformer serves exclusively as a multimodal fusion module, as opposed to functioning as a feature extraction backbone.

2.4. Video-based Person reID

Unlike image-based reID, video-based reID uses a sequence of images to represent an individual, which provides temporal information and a richer set of appearance details Ye et al. [2021b]. Current methods in this domain primarily utilize techniques such as RNNs Liu et al. [2019], Xu et al. [2017], temporal pooling strategies such as average and weighted pooling Xu et al. [2017], optical flow analysis Liu et al. [2019], and 3D convolution Chen et al. [2018], among others. For instance, the work Xu et al. [2017] integrated original person images with the corresponding optical flow into the network input to maintain motion consistency across different time steps. They subsequently utilized the

features extracted from the CNN-RNN to compute attention vectors, which allow selection of informative frames from the sequence. Another approach is the Spatial and Temporal Memory Networks Eom et al. [2021] that store features for spatial distractors, which are frequently encountered across video frames. Additionally, it optimizes attention for typical temporal patterns. This dual approach facilitates a more comprehensive understanding of the temporal and spatial information in video-based person reID tasks.

STRF Aich et al. [2021] introduced a flexible feature processing module capable of being integrated into any 3D convolutional block for reID tasks. This innovative module adeptly captures complementary person-specific appearance and motion information, thereby enhancing the reID process. Further advancements in this field have seen the adoption of 3D graph convolution for video-based reID. For instance, the work Liu et al. [2019] leveraged topology enhanced by context information to construct a graph, which effectively encodes both contextual and physical attributes related to the human body and offers a comprehensive understanding of the subject's identity. While many studies have devoted attention to image-based cross-modal reID or video-based RGB reID, the emergence of the HITSZ-VCM dataset has paved the way for research in video-based visible-infrared person reID comprised of 24-hour surveillance scenarios Lin et al. [2022]. In contrast to these existing methods, our proposed approach leverages a Cross-Modal Fusion Transformer to capture cross-modality interaction and propagation across multi-scales. By integrating this transformer with dual-stream sequence CNN, we can extract richer and more complementary information from both modalities, resulting in improved person reID performance.

3. Methodology

In the previous studies Lu et al. [2023], Lin et al. [2022], Chen et al. [2022], thermal and appearance characteristics were used independently, yet each type of feature carries its inherent limitations. For instance, thermal features disregard an individual's attire and heat signatures, and this information cannot be drawn from an RGB image. Conversely, physical features can rapidly become indistinguishable and do not encapsulate an individual's unique gait pattern Sun et al. [2020]. n this paper, we have leveraged complementary thermal features derived from the infrared input and appearance features obtained from the corresponding visible frames. We introduce the dual-stream 3D feature extraction designed for visible-infrared feature extraction from the video input, followed by the proposition of the CMFT method to derive effective multi-modal representations. Fig 3 illustrates how our model fuses CMFT with 3D temporal shift CNNs to ensure precise person reidentification. Subsequently, we recommend a step-by-step learning strategy to manage the substantial modality gap. Moreover, we utilize a modality-shared refinement loss (MSFL) to further enhance the complementary features. Finally, we introduce a Modality Discriminative Loss (MDL) to further augment the discriminatory power of the reliable modality-shared features.

3.1. Sequence Generation

Within a sequence of frames, various individual instances and background regions coexist. To create a region-of-interest (ROI) hypothesis that exclusively contains a person's body region, we merge the capabilities of YOLOv7 Wang et al. [2023] an end-to-end detection model with the tracking abilities of StrongSORT Du et al. [2023]. YOLOv7 Wang et al. [2023] is a considerable advancement from YOLOv4 Bochkovskiy et al. [2020] marked by a suite of modifications to its architecture. It integrates compound scaling, extends the efficient layer aggregation network (EELAN), and incorporates strategies such as premeditated and reparameterized convolution, granularity for auxiliary loss, and precision for lead loss. A vital part of YOLOv7, the EELAN, enhances the model's learning capability while preserving the original gradient path, made possible by an expand-shuffle, and merge-cardinality techniques. The compound scaling mechanism, introduced to YOLOv7, is designed to retain the model's initial level characteristics and optimal design. Furthermore, enhancement demonstrates the application of model reparameterization and dynamic label assignment for network optimization, addressing prevalent issues and boosting the overall performance of the detection framework.

During the tracking phase, we utilize StrongSORT Du et al. [2023], an advanced iteration of the DeepSORT [53] algorithm. DeepSORT augments the capabilities of the SORT Wojke et al. [2017] algorithm, which conducts object tracking by utilizing the Kalman filter for motion prediction and the Hungarian algorithm for identity association between successive frames. Nonetheless, SORT faces challenges when dealing with occluded objects, often leading to inaccurate associations and identity shifts. To mitigate these issues, DeepSORT Wojke et al. [2017] incorporates a pre-trained CNN to capture the appearance features of objects across the most recent 100 frames, thereby alleviating identity shifts triggered by occlusions. Additionally, DeepSORT introduces cascade matching and trajectory confirmation mechanisms, improving trajectory prediction and object matching in the present frame. StrongSORT, an enhanced version of DeepSORT, introduces two additional lightweight algorithms: AFLink and GSI. AFLink couples short trajectories with completed trajectories utilizing a fully-connected model, bypassing reliance on appearance information. This strategy enhances trajectory completeness while reducing fragmentation. GSI, on the other hand, bolsters detection reliability by emulating nonlinear motion patterns and employing Gaussian regression for precise object positioning. Notably, GSI integrates the motion information of detected objects during the regression process, fostering more resilient and precise trajectory estimation. This combination of YOLOv7 and StrongSORT performs precise person detection and tracking, thereby enabling the formulation of

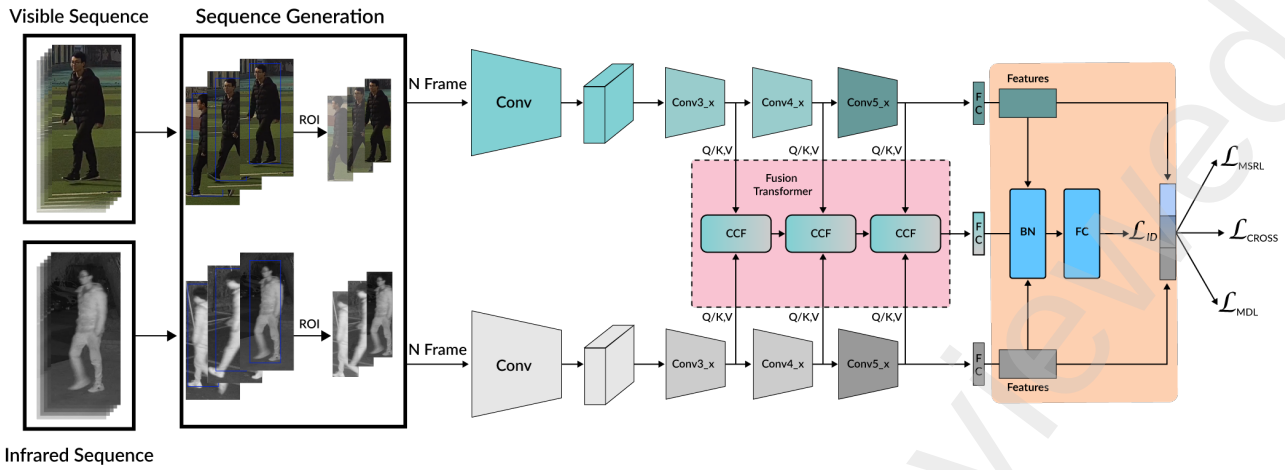

Figure 3: The architecture of our proposed CMFT for visible-infrared video as cross-modal inputs for person reID. CMFT is consisted of 4 main components: 1) Sequence generation which uses Yolov7 for detection and StrongSort for tracking the instance, 2) dual-stream feature extractor which receives the person RoIs as input sequence, 3) CCF for progressive cross-modality feature interaction and fusion at the different Conv layers of dual-stream feature extractor, 4) Progressive learning module which is comprised of two main stages: 1) First stage is supervised by L_{ID} and L_{Cross}, to extract features that don't depend on the specific modality. 2) Second stage is supervised by the L_{MSRL} and L_{MDL} to further develop the features that are shared between modalities.

credible ROI sequence generation for subsequent processing stages.

3.2. Dual-Stream Feature Extractor

We employ a dual-stream architecture to extract the features from RGB and Infrared modality video streams. This dual-stream feature extractor consists of two identical, independent sequence CNNs, which concurrently extract features from both visible and infrared modalities, as depicted in Figure 3. Given the inherent limitations of 2D CNNs in capturing the temporal dynamics in human sequences and the high computational burden imposed by 3D CNNs led to the use of Temporal Shift Module (TSM) Lin et al. [2019] as our base feature extractor. TSM effectively generates unimodal spatio-temporal representation, exhibiting the same spatio-temporal modelling capacity as 3D CNN without any supplementary computational overhead or extra parameters. As depicted in Figure 3, we initially select N pairs of frames $((X_1^{RGB}, X_1^{IR}), (X_2^{RGB}, X_2^{IR}), , , , (X_N^{RGB}, X_N^{IR}))$ from the given visible-infrared sequence. Subsequently, the dual-stream TSM calculates the average output logits from N frames of each modality to determine the reID output score. To combine the complementary features from both modalities, we employ score fusion to acquire the final output prediction from the dual-stream CNNs. Throughout the study, we utilize ResNet-50 He et al. [2016] as a feature extractor of TSM.

3.3. Cross-Modal Fusion Transformer

As shown in Figure 3, the proposed cross-modal cross-scale fusion transformer (CMFT) consists of a sequence of CCFs across various feature scales. These CCFs establish associations between neighbouring feature integration stages. The intermediary features at several stages from the dual-stream TSM are fed into the CCFs to facilitate multimodal interaction and fusion. We explain the vanilla attention mechanism and our CCFs in detail in the following sub-section.

3.3.1. Attention Mechanism

The Self-Attention (SA), a base component of the vanilla Transformer Bertasius et al. [2021], is also a vital component in our CMFT especially multi-head self-attention (MHSA). An input sequence, denoted by M d_z-dimensional vectors $Z = (z_1, z_2, z_3, ... , z_M) \in R^{Mxd}$ space, is provided. The input vectors of each head h, designated as query Q^h, key K^h, and value V^h, are determined via learned linear projections:

$$(Q^h, K^h, V^h) = (ZW_Q^h, ZW_K^h, ZW_V^h) \tag{1}$$

The parameter matrices W_Q^h, W_K^h, and $W_V^h \in R^{d_z x d_h}$, where in d_h is designated as d_z/H, are employed to compute query-key-value vectors. These vectors are subsequently transformed to an attended vector of head h_i via the scaled dot-product operation:

$$h_i = Softmax\left(\frac{Q^h(K^h)^N}{\sqrt{d_h}}\right)V^h \tag{2}$$

The results of all heads are collated and subjected to a further linear transformation leveraging weights $W^0 \in R^{(d_h.H)xd_z}$. Consequently, we attain the ultimate attended sequence $\tilde{Z} = \tilde{z_1}, \tilde{z_2}, ... , \tilde{z_N} \in R^{Mxd_z}$, derived from the multi-head attention mechanism.

$$Z = A(Q, K, V) = [h_1, h_2, ..., h_H]W^O \tag{3}$$

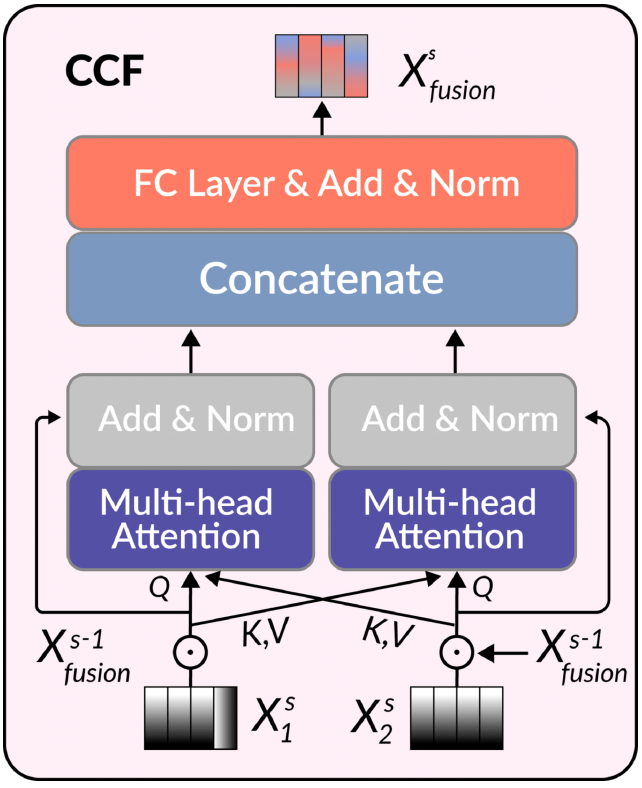

Figure 4: Design of the CCF module.

3.3.2. *Cross-Modal Cross-Scale Fusion*

While RGB and thermal features offer substantial complementary information, the RGB component captures the colourful aesthetic of the image while the thermal component reflects temperature gradients, each contributing valuable insight from their respective domains. However, capturing the relationship between these two distinct feature sets is inherently challenging, primarily due to their different natures and information presentation modalities. To navigate this problem, we introduce CCF to probe into cross-modal interactions at multiple scales using cross-attention. Despite the application of such a cross-attention in other multi-modality problems Ma et al. [2021b], Tang et al. [2023], such as text-to-image retrieval Xu et al. [2023], He et al. [2016], text-to-reID Ma et al. [2021a] and visual question answering Zhang et al. [2021b], our approach exhibits two distinct characteristics. Firstly, the one-to-one alignment of RGB frames and thermal maps, both of which are image data, enables a more straightforward and efficient feature-level interaction and fusion. This contrasts the prevalent misalignment between textual and visual inputs in vision-language models, which can complicate the computational mechanics of the attention mechanism. Secondly, unlike typical image-text-related tasks which mostly concentrate on inferring relationships between preprocessed features using a cross-attention mechanism, our CCF is designed to concurrently facilitate both cross-modal and cross-scale representation learning.

Figure 4 illustrates the CCF architecture comprises of a pair of MHSA coupled with a fully connected (FC) layer. The main function of the CCF is to facilitate the fusion of features extracted from adjacent stages of a dual-stream CNN, which in turn enables a fusion process that operates on multiple scales and modes. Specifically, features $z^s_1 \in R^{WH\times d}$ and $z^s_2 \in R^{WH\times d}$, extracted from the dual-stream feature extractor at stage s, are introduced into separate MHSA layers as distinct roles (labelled, Q, K or V) within the cross-attention (CA) Wang et al. [2022a]. This approach enables the CCF to focus on the relationships between the two modalities, learning how they complement each other Tang et al. [2023], Yang et al. [2022b]. By concatenating the outputs from the two CA layers with the respective final dimension of CV, and taking this as input to the FC layer, the combined multimodal feature $z^s_{fusion} \in R^{WH\times 2d}$ is obtained. Feature recalibration is carried out using multiplication operations to leverage multi-scale features and establish connections between different stages, as affirmed by its effective implementation in Das et al. [2020] for modulating multi-modality learning. However, due to computational limitations, we avoid executing complex operations or introducing additional layers. Instead, we multiply the multimodal feature $z^{s-1}_{fusion} \in R^{WH\times 2d}$ from the preceding stage's CCF by the unimodal features of the current stage prior to the cross-attention operation. This approach allows us to introduce information from lower-level fusion, enabling the model to carry out complex interactions and integrations across scales and modes. Finally, to enhance optimization, we implement a residual connection He et al. [2016] followed by layer normalization (LN) Ba et al. [2016]. This combination of methods allows us to effectively fuse multimodal information at multiple scales, improving the performance of our model. This entire process can be formulated as follows:

$$z^s_{fusion} = LN(FC([z^s_{i,j}, z^s_{i,j}]) + [z^s_{i,j}, z^s_{i,j}]) \quad (4)$$

$$z^s_{i,j} = LN(A((Q = \tilde{Z}^s_i, K = \tilde{Z}^s_j, V = \tilde{Z}^s_j) + \tilde{Z}^s_i)) \quad (5)$$

$$\tilde{z}^s_i = Z^s_i \odot Z^{s-1}_{fusion}) \quad (6)$$

For i, j = 1, 2, where i ≠ j, and ⊙ denoting the element-wise product function, we can see that the positioning of the initial CCF plays a significant role in determining the complexity of information integrated into the model. By structuring the CCFs in a cascade arrangement across various feature levels, the model is able to systematically enrich multimodal features and propagate them across different scales. It's worth noting that, in the final CCF, we utilize a pair of FC layers prior to the MHSA operation (Equation 5) in order to reduce the embedding dimensions d and the number of parameters. We set the output dimension for these FC layers to 1024 to balance computational efficiency and

model complexity. The choice of the specific location for the first CCF and the number of MHSA layers L within the CCFs are crucial design decisions that can impact the performance of the model. We explore these aspects in greater detail in Section 4.3, where we discuss the impact of these design choices on the overall performance of the model and provide insights into the optimal configuration.

3.4. Progressive Learning

3.4.1. Cross-Modal Loss

As illustrated in Figure 3, the model training process is comprised of two stages. During the initial stage, the framework utilizes Z^{RGB}, Z^{IR}, Z^{Fusion} as inputs, independently selecting positive and negative person instances from each modality. By leveraging L_{Cross}, the CMFT primarily focuses on extracting the discriminative patterns across modalities in the second training stage, which effectively mitigates the negative impacts brought about by the substantial disparity between visible and infrared modalities. With L_{Cross}, positive and negative samples are based on the distance of the cross-modality fused feature. This approach allows the model to retain raw image information, benefiting from modality-specific and cross-modality complementary information.

$$L^{RGB} = \sum_{i=1}^{N} [max_{y_i=y_j} D(Z_i^{RGB}, Z_j^{RGB}) - min_{y_i \neq y_k} D(Z_i^{RGB}, Z_k^{RGB}) + m] \quad (7)$$

$$L^{IR} = \sum_{i=1}^{N} [\max_{y_i=y_j} D(Z_i^{IR}, Z_j^{IR}) - \min_{y_i \neq y_j} D(Z_i^{IR}, Z_k^{IR}) + m] \quad (8)$$

$$L^{Fusion} = \sum_{i=1}^{N} \left[\max_{y_i=y_k} D(Z_i^{Fusion}, Z_k^{Fusion}) - \min_{y_i \neq y_k} D(Z_i^{Fusion}, Z_k^{Fusion}) + m \right] \quad (9)$$

$$L^{Cross} = L^{RGB} + L^{IR} + L^{Fusion} \quad (10)$$

Here, D(.) stands for a distance metric, y_i represents the identity label of the i-th image, $[z]_+ = max(z, 0)$, and m and m denotes a margin.

3.4.2. Modality-Shared Refinement Loss

In real-world scenarios, visible and infrared images often exhibit substantial modality disparities. Consequently, it becomes vital to extract features that are invariant across modalities. As illustrated in Figure 5, the red back side is visible solely in the RGB modality. Overemphasis on these features may lead to inaccuracies during cross-modality retrieval. Hence, we incorporate the MSRL to reduce the impact of unstable features that only show up in one modality, encouraging the use of steady features that are consistent across both modalities. To get the complimentary features, we investigate latent information embedded in all samples within a given mini-batch. Formally, we represent the anchor features of the RGB, infrared and fused modalities as z^{RGB}, z^{IR}, and z^{Fusion}, respectively. Initially, we compute its mean distance to from each feature to the other positive samples within the intra-modality and cross-modality. This is represented as follows:

Figure 5: The proposed MSRL diminish the features that are unreliable and unimodal, while simultaneously boosting the features that are shared across different modalities.

$$D^{RGB} = \frac{1}{K} \sum_{i=1, i \neq j}^{N} D(Z_j^{RGB}, Z_i^{RGB}) \quad (11)$$

$$D^{IR} = \frac{1}{K} \sum_{i=1, i \neq j}^{N} D(Z_j^{IR}, Z_i^{IR}) \quad (12)$$

$$D^{Cross} = \frac{1}{K} \sum_{i=1, i \neq j}^{N} D(Z_j^{Fusion}, Z_i^{Fusion}) \quad (13)$$

Then, the L_{MSRL} is defined as:

$$L_{MSRL} = \frac{1}{2PN} \sum_{P=1}^{2N} \left[\sum_{j=1}^{2N} D(D_j^{RGB}, D_j^{cross})^2 \right.$$

$$+\sum_{j=1}^{2N} D(D_j^{IR}, D_j^{cross})^2\Big] \quad (14)$$

As reflected in Equation 7, L_{MSRL} punishes the disparity between intra-modality distance D^{RGB} and D^{IR}, and cross-modality distance D^{cross}. If distinctive features emerge solely within one modality, then the discrepancy between D^{RGB}-D^{IR} and D^{cross} escalates, a deviation detected by L_{MSRL}. During the bi-directional optimization process involving D^{RGB}- D^{IR} and D^{cross}, the less reliable features within a single modality are suppressed, while the more reliable features appearing across both modalities are amplified.

3.4.3. Modality Discriminative Loss

Similar to the challenges in RGB-RGB person reID, a single individual may encounter substantial intra-class variations attributable to typical changes in aspects such as posture, perspective, lighting, etc. These variations significantly amplify the complexity of aligning features across different modalities. To tackle this issue, similar to Zheng et al. [2023], we introduce a Modality Discriminative Loss (MDL) which leverages the relationships between central instances and augments the discriminative capacity of reliable features shared across modalities. We calculate the discriminative representation from both modalities as:

$$c_{yi} = \frac{1}{2N}(\sum_{j=1}^{N} Z_j^{RGB} + \sum_{k=1}^{N} Z_k^{IR}) \quad (15)$$

Here c_{yi} signifies the central feature of the y_i^{th} identity. Subsequently, we compute the average distance from c_{yi} to all other negative samples, which serves as a dynamic margin and can be symbolically represented as follows:

$$d_{yi}^{n}eg = \frac{1}{2N}(P-1)\sum_{y_i \neq y_j} ||Z_j - c_{y_i}||_2 \quad (16)$$

Finally, the L_{MDL} is defined as:

$$L_{MDL} = \frac{\sum_{i=1}^{P} mean_{y_j=y_i}||Z_j - c_{y_i}||_2}{\sum_{i=1}^{P} mean_{f_k=c_{y_i}}||Z_k - c_{y_i}||_2} \quad (17)$$

Through the minimization of Equation 10, enhancements can be achieved in both intra-class compactness and inter-class discrimination. The implementation of L_{MDL} has two primary benefits: 1) It can capture modality-specific information and extract better complementary relationships. 2) Dynamic sampling via $d_{y_i}^{neg}$ allows for an effective focus on relatively challenging examples.

3.4.4. Objective Function

To train the model, we employ a compound loss function within our progressive learning infrastructure Zhang et al. [2023]. In the initial phase, we leverage the identity loss L_{ID} Zheng et al. [2017] and L_{Cross} to facilitate the learning of modality-independent features:

$$L_1 = L_{Cross} + L_{ID} \quad (18)$$

In the subsequent stage, we go deeper into the extraction of dependable modality-shared features using L_{MSRL}, and the discriminatory power is augmented with the aid of L_{MDL}. The loss function at this stage is computed as follows:

$$L_2 = L_{ID} + \lambda_1 L_{MSRL} + \lambda_2 L_{MDL} \quad (19)$$

Here, the parameters λ_1 and λ_2 are used control the impact of L_{MSRL} and L_{MDL}, respectively.

4. Experimental setup

In this section, we have reported the used datasets, experimental setup, and evaluation of the different components of the proposed CMFT through a comprehensive ablation and performance comparison with different benchmark reID methods.

4.1. Dataset

HITSZ-VCM Lin et al. [2022] is a comprehensive visible-infrared person reID dataset-based video. It is compiled from a vast array of images/frames, collected by deploying 12 high-definition cameras each offering a resolution of 3840 × 2160. Thanks to advancements in surveillance technologies, these cameras can flawlessly switch between capturing RGB and IR images based on the prevailing light conditions, thereby guaranteeing that each individual is recorded in both RGB and IR modes. Encompassing 927 unique identities, the HITSZ-VCM dataset is bigger enough for network training. At a frame rate of 25 FPS, the first frame out of every set of 5 frames is selected to compile the final dataset. A sequence of 24 consecutive images is viewed as a tracklet, which symbolizes an individual during a specific time period. There may be fewer than 24 frames in the final tracklet. This configuration ultimately leads to 251,450 RGB and 211,800 IR images approximately, systematically arranged into 11,785 and 10,078 tracklets, respectively. The dataset allows for dynamic adjustment of the number of frames per tracklet, offering greater flexibility than many other video-based datasets currently available. The HITSZ-VCM dataset, with its 12 recording cameras, ensures that most identities are recorded by 3 RGB and 3 IR cameras, with no overlap. The dataset captures a diverse range of environments, encompassing 7 outdoor, 3 indoor, and 2 passageway settings. These include commonplace locations such as offices, coffee shops, corridors, playgrounds, and gardens. Each individual is captured from various angles under each camera, thereby yielding a rich and comprehensive dataset of appearances. In addition, the dataset incorporates complex scenarios such as changes in lighting belongings changes, obstructions, and shifts in viewpoints.

The SYSU-MM01 Wu et al. [2017]is a challenging benchmark for cross-modal reID, specifically focusing on infrared-visible cross-modal scenarios. It comprises images captured in both indoor and outdoor settings, featuring 419 individuals. The dataset includes images from four visible and two near-infrared cameras, resulting in approximately 30,000 RGB and 15,800 infrared images from 491 distinct

individuals. Each individual is captured by at least one visible and one near-infrared camera, enabling cross-modal comparisons. The dataset offers two testing configurations: indoor-search and all-search Xiao et al. [2017], designed for evaluating RGB-IR reID methods. All-search includes images from both indoor and outdoor RGB cameras, providing a versatile approach suitable for diverse environmental conditions. Conversely, indoor-search mode restricts the gallery. The train set consists of 395 identities with around 22,258 RGB and around 11,900 infrared images. The testing set comprises 96 identities and includes a query set and a gallery set. The query set contains 3,803 infrared images, and the gallery set consists of 301 randomly selected visible images.

RegDB Park et al. [2021] is a dual-camera dataset designed for cross-modal reID research. The dataset consists of a total of 8,240 images, capturing 412 individuals. Each individual is represented by 10 RGB and 10 infrared images. Dataset assigns 4,120 images (10 images per person) of 206 individuals for training purposes. The remaining 4,120 images featuring the other 206 individuals are designated for inference. For inference, a gallery set and a probe set are formed by separating images from the RGB-IR. Specifically, the gallery set is composed of images from one modality (either RGB or IR), while the probe set consists of images from the other modality. This process is repeated for 10 trials, with different combinations of gallery and probe sets, to ensure a statistically robust evaluation. RegDB is designed to simulate real-world scenarios by employing dual camera systems to capture RGB and thermal images of individuals. The dataset facilitates the evaluation of cross-modal person reID methods and allows for the assessment of the model's performance in scenarios involving different modalities. The availability of multiple trials and the separation of gallery and probe sets based on different modalities provide a rigorous evaluation setup, enabling researchers to obtain reliable and meaningful results in their cross-modal person reID studies.

4.2. Training and Evaluation Settings

Our proposed CMFT uses a set of matching RGB-Infrared frames, denoted as $\{X^{RGB}, X^{IR}\}$, as its input. The number of frames used is set at {6, 6} and {12, 12}, which is specifically designed for detailed studies and comparison with top methods. The experiment is run on a high-performance Nvidia RTX3090 GPU. For improving the data, we use the techniques outlined in Lin et al. [2022]. As part of this improvement process, all images of people are resized to a size of 256×128. Other improvement techniques, such as turning the image horizontally and randomly erasing parts, are also used. For infrared images, we add more techniques like colour jittering and Gaussian blurring. The main part of our model uses parameters that have been pretrained on the ImageNet dataset Deng et al. [2009]. To finetune the model, we use minibatch stochastic gradient descent with a momentum factor of 0.9, over a total of 60 epochs. The batch size is fixed at 64 for this process. The initial learning rate is set at 0.1 and is slowly reduced by a factor of 0.01 every 10 epochs. The weight decay parameter is fixed at 1e-2 for this particular study. By default, we use 8 heads for each instance of MHSA. The model is trained over 24 epochs on the HITSZ-VCM dataset and is increased to 32 epochs for the SYSU-MM01. For both datasets, during the warm phase of training, the epoch is set to 8, while the loss balancing parameters λ_1 and λ_2 are set to 0.5. After experimentation with different values, 0.3 and 0.7 come out as the optimal values for λ_1 and λ_2 The margin parameter m is fixed at 0.1. During the testing phase, we use the 1024- and 512-dimensional features after the batch normalization (BN) layer. Our testing method follows the commonly used testing setting as explained in Lin et al. [2022], Lu et al. [2023], which includes cropping and flipping techniques.

In order to effectively manage GPU computational and memory limitations, we adopt the Swin-S Liu et al. [2022] structure as the backbone for constructing a transformer-based network. From each visible-infrared sequence, we uniformly sample a clip composed of 16 frames with a stride of 2. This sampled clip then serves as the input for the network. During the training phase, we make use of a cosine decay learning rate scheduler, which includes 3 epochs of linear warmup, and run the training operation for a total of 60 epochs. We use a batch size of 8 for this process. The remaining training configurations closely follow the guidelines as described in Liu et al. [2022]. For the inference phase, we implement a 3-crop strategy with a single clip, adhering to the approach recommended in Liu et al. [2022], Li et al. [2022]. As a result, the final prediction is calculated by averaging the scores derived from 31 different views. This asserts that the final prediction is robust and reflects the comprehensive analysis carried out during the inference phase.

In this study, we perform cross-camera and cross-modal retrieval tasks, similar to previous works Fu et al. [2021], Ye et al. [2020b], Jiang et al. [2022], Chen et al. [2022], Lin et al. [2022], Lu et al. [2023]. This means that the query and gallery samples are captured using different cameras and modalities. To maintain consistency with the training data, we adopt a video-video matching approach for the probe-gallery pattern. Specifically, for the HITSZ-VCM dataset, we employ two retrieval modes: infrared-visible and visible-infrared, to enable a comprehensive evaluation. In the infrared-visible retrieval mode, the query set comprises 5,159 tracklets, while the gallery set consists of 5,643 tracklets. Conversely, in the visible-infrared mode, the query set contains 5,643 tracklets, and the gallery set consists of 5,159 tracklets. It is worth noting that, during our implementation, we exclude tracklets that are too short, specifically those containing less than 12 images, to ensure the reliability of the results.

Our performance evaluation metrics include Rank-1, Rank-5, Rank-10, Rank-20, mAP, and Mean Inverse Negative Penalty (mINP). These metrics are commonly used and accepted in the field for their capacity to provide robust and comprehensive evaluations. Similar to a number of

Table 1
Evaluation of different person detectors on HITSZ-VCM (Visible part) dataset.

Matrix	Faster R-CNN		YOLOv5		YOLOv7	
	Day-Detection	Night-Detection	Day-Detection	Night-Detection	Day-Detection	Night-Detection
Precision	0.893	0.861	0.909	0.880	0.935	0.913
Recall	0.917	0.874	0.916	0.906	0.941	0.936
F1-Scores	0.905	0.867	0.912	0.893	0.937	0.924
Accuracy	0.946	0.926	0.951	0.948	0.973	0.964
FPS	14	14	26	26	29	29

existing methods Ye et al. [2020b], Li et al. [2023], Ye et al. [2020b], we compute the distance scores between the query features and the gallery features in order to perform a ranking operation. This involves comparing each query feature with all the features in the gallery and determining the distance between them. The rankings are then established based on these distance scores. During the testing phase, we utilize cosine similarity as the primary measurement for comparing distances. Cosine similarity is particularly well-suited to high-dimensional data as it assesses the cosine of the angle between two vectors, providing a measure of their orientation, rather than their absolute distance.

4.3. Ablation

we conduct the ablation study of the different components of the proposed CMFT on the HITSZ-VCM. In this particular subsection, we undertake an empirical analysis to analyze the significance and impact of our CMFT on the HITSZ-VCM dataset. This involves a close investigation of the efficacy of the various strategies that have been integrated into the proposed CMFT method. The objective of this analysis is to not only validate the effectiveness of the proposed model but also to understand the contribution of each component. The default setting for the number of MHSA layers incorporated in each branch of CCF is set to 1 unless explicitly specified otherwise. Loss balancing parameters λ_1 and λ_2 are set to 0.3 and 0.7 thoughtout the ablation.

4.3.1. Performance comparison of Sequence Generation

Real-world reID requires detecting multiple persons to form the sequence of each person. In this section, we evaluate the performance of different state-of-the-art object detectors: Such as Faster R-CNN Ren et al. [2015], Yolov5 Zhu et al. [2021b], and Yolov7 Wang et al. [2023]. Faster R-CNN [60] is known for higher performance, but lower FPS rate, while Yolov5 and Yolov7 require very less computational resources as compared to Faster R-CNN with comparative performance. To select the detector with a better tradeoff between efficiency and accuracy, both the Faster R-CNN and the YOLOV5 detection models are used to train on 800-person images and test on 200-person images. YOLOv5 divides the CNN feature map into a fixed-sized grid and predicts the fixed number of bounding boxes with its score on every grid for better recall and accuracy.

$$Precision = \frac{Positive\ Predicted\ Correctly}{All\ positive\ Predictions} \tag{20}$$

$$Recall = \frac{Positive\ predicted\ correctly}{All\ positive\ Observations} \tag{21}$$

$$F1 - score = \frac{2(Precision * Recall)}{(Precision + Recall)} \tag{22}$$

The table provided indicates a comparative performance analysis of various object detectors, Faster R-CNN, YOLOv5, and YOLOv7, in different conditions: day and night detection. According to the evaluation's findings, the Faster R-CNN model performs similarly to the YOLOV5 Zhu et al. [2021b] model but requires higher computational cost. However, In contrast to Faster R-CNN detections, high-frame rate YOLOV5 detection for human tracking is supplemented in this study to improve performance. The chosen real-time image detector is YOLOV5, which has an inference rate ten times higher than Faster RCNN. The results of both detectors are displayed in Table 1 below.

The Faster R-CNN model achieves a precision of 0.893 during daytime detection, which drops slightly to 0.861 in nighttime scenarios. The F1-scores for this model come out to be 0.905 and 0.867 for day and night detections, while the accuracy scores are noted at 0.946 and 0.926, respectively. Comparatively, YOLOv5 sees an improvement of roughly 1.8%, 0.1%, 0.8%, and 0.5% in the precision, recall, F1-score, and accuracy presenting an improvement of roughly 1.8%, 0.1%, 0.8%, and 0.5% respectively, when compared to Faster R-CNN. The improvement in nighttime detection is even more pronounced with increases of approximately 2.2% for precision, 3.7% for recall, 3% for F1-score, and 2.4% for accuracy. An important highlight is that YOLOv5 achieves a higher FPS rate of 26, as compared to 14 FPS in Faster R-CNN, indicating a faster processing time. However, YOLOv7 stands out by delivering the best performance among the three models. For daytime detection, it outperforms Faster R-CNN by approximately 2.8% for precision, 2.6% for recall, 2.8% for F1-score, and 2.3% for accuracy. In comparison to YOLOv5, the improvements are around 0.9% for precision, 2.7% for recall, 2.2% for F1-score, and 2.3% for accuracy. In the context of nighttime detection, YOLOv7 shows about a 6% improvement for precision, 7.1% for recall,

Table 2
Evaluation of cross-modality fusion techniques for the dual-stream feature extractor on HITSZ-VCM dataset

Techniques	Rank-1	Rank-5	Rank-10	mAP	mINP	Parameters	FLOPs
Score average	58.52	76.01	79.94	39.13	31.67	-	-
Score Multiplication	60.31	76.24	81.51	40.70	34.38	-	-
Feature concatenation	61.70	78.07	82.69	41.83	35.46	-	-
$CMFT_{S6F}$	63.24	79.58	85.03	47.62	39.45	8.39M	1.24G
$CMFT_{D6F}$	69.88	84.71	89.65	55.68	46.81	17.29M	1.63G

6.6% for F1-score, and 4.1% for accuracy compared to Faster R-CNN. When compared to YOLOv5, the improvements are 3.7% for precision, 3.3% for recall, 3.5% for F1-score, and 1.7% for accuracy. Additionally, YOLOv7 operates at the highest FPS of 29, implying the fastest processing speed.

4.3.2. Comparison on Fusion Techniques

This section provides a comparison of the performances of three different modality feature fusion methods: score, feature concatenation, and the proposed CCF in the CMFT method. Each of these methods utilizes the same dual-stream feature extractor and parameter settings to maintain uniformity in comparison. Score fusion represents a method that combines the output vectors of the dual-stream feature extractor. This combination is achieved using either averaging or element-wise multiplication, thereby yielding a singular score from multiple inputs. On the other hand, feature concatenation fusion integrates the unimodal information derived from the two streams. This integration is achieved by feeding the concatenated feature into an FC layer, effectively merging the features into a unified representation. We examine two design choices of CCF, One variation includes two cross-modal interaction sub-branches, while the other variation includes only a single sub-branch. This comparison aids in understanding the impact of the number of sub-branches on the CCF module within the CMFT method. Specifically, $CMFT_{D6F}$ refers to fusing features at Conv6 using a single-branch CCF, providing only single-scale information. On the other hand, $CMFT_{D6F}$ is identical to $CMFT_{S6F}$, except for the inclusion of a CCF with two branches (as depicted in Figure 3).

As shown in Table 2, score average displays a Rank-1 score of 58.52, mAP value of 39.13, and mINP score of 31.67, while score multiplication method shows a slight improvement, achieving a Rank-1 score of 60.31 (3% increase), mAP score of 40.70 (4% increase), and mINP score of 34.38 (8.5% increase). Feature concatenation shows an approximately 2.3% increase in Rank-1 from the score multiplication, moreover, it also sees a 2.8% increase in mAP and a 3.1% increase in mINP. Moving forward, $CMFT_{S6F}$ provides significant improvements across all metrics, with a Rank-1 score of 63.24, mAP of 47.62, and mINP of 39.45. Compared to the feature concatenation, these improvements represent an approximate increase of 2.5%, 13.9%, and 11.3% in Rank-1, mAP, and mINP scores respectively. $CMFT_{S6F}$ has 8.39M for parameters and 1.24G FLOPs. In comparison to the $CMFT_{S6F}$, $CMFT_{D6F}$ yields a Rank-1 score of 69.88, mAP of 55.68, and mINP of 46.81 and indicates 10.5%, 17%, and 18.7% performance improvement for Rank-1, mAP, and mINP, respectively. However, this superior performance does come at the cost of a larger model due to 17.29M parameters and greater computational complexity shown by 1.63G FLOPs. The performance of $CMFT_{D6F}$ surpasses that of other methods, thereby emphasizing the importance of utilizing a dual-branch cross-modal interaction. This configuration facilitates a reciprocal interaction of features between the modal branches. As a result, it facilitates a more comprehensive learning experience by enabling the extraction of richer complementary information from each respective branch. By encouraging these symmetrical feature interactions, the two-branch cross-modal interaction strategy enhances the CMFT's ability to identify and leverage the unique and shared characteristics present in each modality. This ability ultimately leads to a more robust and effective model, as demonstrated by the superior performance of $CMFT_{D6F}$.

4.3.3. Abalation of Multi-scale Fusion

In this section, we investigate the effectiveness of multi-scale interaction for multimodal reID by CCF modules at various scales. $CMFT_{Si}$ denotes the fusion of features starting from the i-th feature extraction stage of the TSM as illustrated in Figure 3. As shown in Table 3, $CMFT_{S6}$ presents a 70.54 rank-1 score, mAP value of 53.71, and an mINP score of 45.20. $CMFT_{S5}$ witness a 0.9% decrease in the rank-1 as compared to $CMFT_{S6}$. However, the mAP and mINP shows a 3.7% and 3.6%, respectively. $CMFT_{S4}$ fusion experiences a drop across all metrics compared to the $CMFT_{S5}$ level. It records a Rank-1 score of 68.27 (a 2.3% decrease), an mAP score of 54.04 (a 2.9% decrease), and an mINP score of 43.52 (a 7% decrease). $CMFT_{S3}$ exhibits a continued reduction in performance and see a 2.4%, 2.9%, and 1.5% decrease as compared to $CMFT_{S4}$, while $CMFT_{S2}$ experience a 1.6% Rank-1, 1.5% mAP, and 4.3% mINP decrease, respectively, when compared to the $CMFT_{S3}$. Results show that $CMFT_{S5}$ achieves the highest performance. Furthermore, while cross-scale interaction among modalities has been shown to amplify multimodal fusion, the incorporation of additional scales appears to have a negative effect on reID accuracy. Additionally, it places considerable computational demands on the system. We posit that the features derived from different modalities

Table 3
Comparison of multi-scale fusion levels for the dual-stream feature extractor on HITSZ-VCM dataset

Techniques	Rank-1	Rank-5	Rank-10	mAP	mINP	Parameters	FLOPs
$CMFT_{S6}$	70.54	84.34	90.43	53.71	45.20	20.65M	4.27G
$CMFT_{S5}$	69.88	84.71	89.65	55.68	46.81	24.34M	6.06G
$CMFT_{S4}$	68.27	82.38	87.17	54.04	43.52	27.71M	7.71G
$CMFT_{S3}$	66.61	82.95	86.94	52.46	42.85	29.83M	9.97G
$CMFT_{S2}$	65.52	80.14	84.94	51.70	41.02	33.40M	13.5G

Table 4
Effect of different cross-modality loss functions on HITSZ-VCM dataset.

Loss	Rank-1	Rank-5	Rank-10	mAP	mINP
L_{CMT} Jiang et al. [2022]	59.45	79.19	82.50	48.43	40.26
WRT Tao et al. [2022]	63.10	82.34	87.76	50.27	41.11
HCT Liu et al. [2020]	67.52	82.43	88.41	53.04	42.30
Our (L_{Cross})	69.88	84.71	89.65	55.68	46.81

at lower levels display significant variability. This diversity presents a substantial challenge when attempting to fuse these features. The heterogeneity among lower-level features complicates the fusion process, potentially contributing to the observed decrease in recognition accuracy when additional scales are introduced. Therefore, a careful balance must be maintained between expanding the model's scope with additional scales and ensuring the effective fusion of the diverse features present across different modalities.

4.3.4. Effectiveness of the Cross-modality Loss

We assess the effectiveness of our cross-modality loss, L_{Cross}, with L_{CMT} Jiang et al. [2022], WRT Tao et al. [2022] and HCT Liu et al. [2020] and comparison results are presented in Table 4. L_{CMT} achieves a 59.45, 48.43 and 40.26, rank-1, mAP and mINP, respectively. WRT notices a 6.15% rank-1, 3.8% mAP and a 2.9% mINP improvement as compared to L_{CMT}, while HCT witness a 7%, 5.5% ad 2.9% increasement in rank-1, mAP and mINP, respectively, as compared to WRT. However, our cross-modality loss, LCross, see 69.88 rank-1, 53.04 mAP and 42.30 mINP which is 3.5%, 5.5% and 2.9 improvement as compared to HCT.

4.3.5. Effects of MSRL and MDL

Table 5 presents a comparison of different settings with MSFL and MDL. $MSFL_{cosine}$ and $MSFL_{euclidean}$ indicate that the model uses Cosine and Euclidean distance, respectively. Results consistently demonstrate that the model with MSRL consistently enhances performance. As indicated in Table 5, the $MSFL_{euclidean}$ model yields the most significant improvements, with enhancements of 5.3% in Rank-1, 6.3% in mAP, and 3.4% in mINP compared to the $MSFL_{cosine}$. Regarding MDL, DCL_{hard} signifies selecting only the closest negative sample for each identity centre. DCL_{all} represents selecting all negative samples for each identity centre, while DCL_{dyn} denotes dynamically

Table 5
Comparison of Results with MSRl and MDL on HITSZ-VCM dataset.

Techniques	Rank-1	Rank-10	mAP	mINP
$MSFL_{cosine}$	61.57	82.72	45.04	40.15
$MSFL_{euclidean}$	64.37	87.11	47.87	41.50
DCL_{hard}	63.98	88.82	49.11	42.39
DCL_{all}	63.61	87.37	48.75	42.46
DCL_{dyn}	65.32	88.61	51.42	43.19
$MSFL_{euclidean}$ + DCL_{dyn}	69.88	89.65	55.68	46.81

selecting negative samples based on Eq. 9. Our dynamic selection method, DCL_{dyn}, demonstrates notably superior results, achieving enhancements of 2.7% in rank-1 score, 5.5% in mAP, and 1.7% in mINP compared to DCL_{all}. When combined with MSRL, $MSFL_{euclidean} + DCL_{dyn}$, the model archives 69.88 rank-1, 55.68 mAP and 46.81 mINP. These findings convincingly illustrate the effectiveness of our MSRL and MDL approaches.

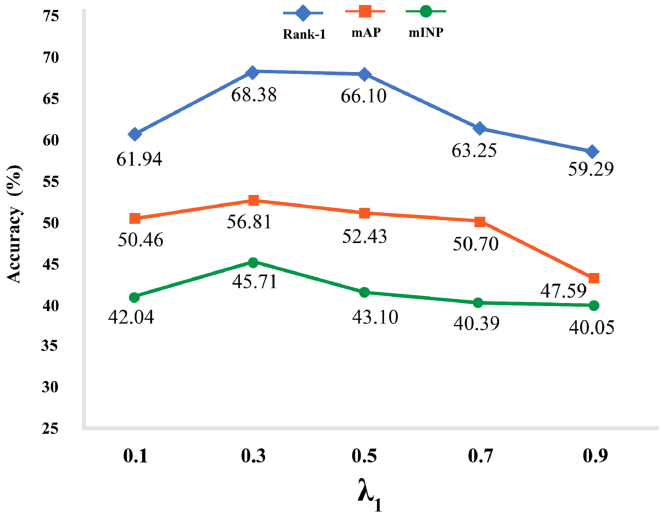

Figure 6: Comparison of weight λ_1 on HITSZ-VCM dataset.

4.3.6. Comparison of Different Loss Weights

As illustrated in Figure 6, the accuracy displays a consistent upward trend until it peaks at $\lambda_1 = 0.3$. Beyond this optimal point, the accuracy begins to decline. This trend suggests that a well-balanced weight, λ_1, allows the

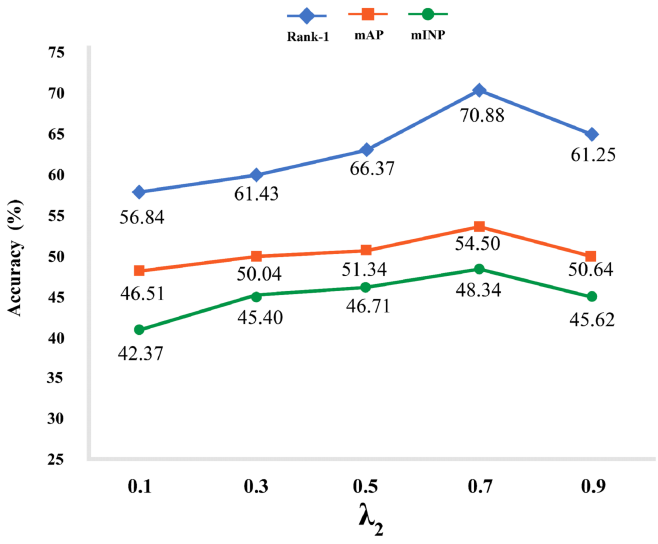

Figure 7: Comparison of weight λ_2 on HITSZ-VCM dataset.

multimodal fusion to equally benefit from both unimodal complementary information. This balance enables the model to effectively leverage the complementary information provided by the two branches. In contrast, the weight λ_2 exhibits an increasing trend until it reaches 0.7 as shown in Figure 7. The results obtained within the range of $\lambda_1 = 0.1$ to $\lambda_1 = 0.9$ provide evidence that the L_{MSRL} contributes less compared to L_{MDL}. When the weight assigned to this branch is decreased, the multimodal features are deprived of valuable information. This lack of information results in outcomes that are less effective than those achieved with a larger weight. These results emphasize the importance of carefully calibrating the weights assigned to different branches of the network. Ensuring the optimal balance of weights allows the model to effectively integrate and leverage the unique and complementary information present in each unimodal branch, ultimately resulting in higher overall accuracy.

4.3.7. Effect of Tracklet Length

Contrasting image data, the more plentiful information offered by video data significantly enhances the performance of person reID task. To validate this claim, we have carried out a series of tests where we have experimented with different numbers of images within a single tracklet which is used as input to our reID model. As shown in Figure 8, a upward trend in the reID performance is observed as the number of images in a tracklet increase which shows that spatio-temporal information provides a boost to reID performance. Analyzing the performance of CMFT with respect to different tracklet lengths, we observe notable trends. As the tracklet length increases, we observe a general increase in performance across all metrics. Specifically, moving from a single image in a tracklet (tracklet length 1) to two images, there's a considerable improvement in Rank-1 accuracy from 38.51% to 43.70% as shown in Figure 9. The trend of improvement continues as we increase the tracklet length, with the highest jump seen between tracklet lengths 3 and 4, where the Rank-1 score moves from 53.22% to 59.57%. This trend of enhancement is prevalent even at Rank-5, Rank-10, mAP, and mINP scores. It is particularly significant at tracklet lengths 5 and 6, where the Rank-1 score improves from 62.46% to 66.50%, and Rank-10 improves from 82.49% to 83.24%. However, as we reach the tracklet length of 7, the increment slows down and the percentage improvement becomes less significant. There is a minor increase in Rank-1 score from 66.50% to 69.88% and in mAP score from 52.77% to 55.68%. Upon moving to a tracklet length of 8, the percentage improvements on all metrics scores become even less significant, almost reaching a plateau. Finally, when the tracklet length is increased to 9, we observe a slight decrease in Rank-5 score from 86.43% to 82.99% and mINP score from 47.54% to 43.63%, indicating that a longer tracklet length does not necessarily equate to a consistent increase in performance. Thus, it seems the optimal tracklet length for achieving high scores with CMFT lies between 6 to 9 images. To minimize computational expenses, we selected n as 7 for all other experiments.

4.4. Performance comparison with different state-of-the-art reID models on HITSZ-VCM dataset.

4.4.1. HITSZ-VCM

We conduct a comparative analysis between our proposed CMFT and several state-of-the-art visible-infrared cross-modal person reID methods, such as DDAG Ye et al. [2020b], LbA Park et al. [2021], MPANet Wu et al. [2021], VSD Tian et al. [2021], CAJL Ye et al. [2021a], MITML Lin et al. [2022], SGIEL Feng et al. [2023], and IBAN Li et al. [2023]. It's important to highlight that these reID methods were primarily developed with image-based datasets in mind. In order to maintain fairness in our comparison, we implement an average pooling to the frame-level features generated by these methods. This approach ensures that the features derived from these methods are processed in a similar manner to our own, allowing for a more direct comparison of performance. For networks utilizing the ResNet50 backbone, such as Long et al. [2022], Chen et al. [2021a], Liu et al. [2019], Lu et al. [2023], we perform average pooling after the backbone, following a similar approach to our baseline. We see significant improvements across all evaluated metrics, both for Infrared-Visible (I-V) and Visible-Infrared (V-I) settings as shown in Table 6. In the I-V setting, the highest Rank-1 score among the other models is held by SGIEL Feng et al. [2023] at 67.65. However, our CMFT outperforms this by achieving a Rank-1 score of 69.88, representing a 3.3% improvement. The same trend is observed across other metrics. The Rank-5 score of our model stands at 84.71, which is an improvement over SGIEL's 80.32, and the Rank-10 score increases to 89.65 from SGIEL's 84.73. Similarly, in terms of mAP, our CMFT sees a significant uplift to 55.68 from SGIEL's 52.30, representing a 6.5% improvement. Looking at the V-I setting, our CMFT continues to outperform other models. The Rank-1 score of our model is 73.37, showing an improvement

Figure 8: The qualitative result visualization of our CMFT results across various traacklet lengths, where n is the number of frames in a trackless. Single image-based reID (with n = 1) fails to produce a satisfying outcome when two identities have similar appearances, while video-based reID demonstrates noticeable improvement in performance, thanks to the inclusion of additional temporal information because of a sequence of the frame.

Table 6
Performance comparison with different SOTA reID models on HITSZ-VCM dataset.

Method	Infrared-Visible					Visible-Infrared				
	Rank-1	Rank-5	Rank-10	Rank-20	mAP	Rank-1	Rank-5	Rank-10	Rank-20	mAP
DDAG Ye et al. [2020b]	54.62	69.79	76.05	81.50	39.26	59.03	74.64	79.53	84.04	41.50
LbA Park et al. [2021]	46.38	65.29	72.23	79.41	30.69	49.30	69.27	75.90	82.21	32.38
MPANet Wu et al. [2021]	46.51	63.07	70.51	77.77	35.26	50.32	67.31	73.56	79.66	37.80
VSD Tian et al. [2021]	54.53	70.01	76.28	82.01	41.18	57.52	73.66	79.38	83.61	43.45
CAJL Ye et al. [2021a]	56.59	73.49	79.52	84.05	41.49	60.13	74.62	79.86	84.53	42.81
MITML Lin et al. [2022]	63.74	76.88	81.72	86.28	45.31	64.54	78.96	82.98	87.10	47.69
SGIEL Feng et al. [2023]	67.65	80.32	84.73	-	52.30	70.23	82.19	86.11	-	52.54
IBAN Li et al. [2023]	65.03	78.34	82.98	87.19	48.77	69.58	81.51	85.43	88.78	50.96
CMFT (ours)	69.88	84.71	89.65	95.52	55.68	73.37	89.51	92.48	97.90	58.44

of 4.5% over the highest Rank-1 score by SGIEL, which stands at 70.23. This improvement trend carries our model reaching 89.51 Rank-5, 92.48 Rank-10, and 97.90 Rank-20 respectively, showcasing noticeable enhancements from the SGIEL's scores of 82.19, 86.11. Furthermore, the mAP for our model is at a robust 58.44, compared to the SGIEL mAP of 52.54, showing a significant increase of 11.2%.

4.4.2. SYSU-MM01

We also compare the performance on the SYSU-MM01 dataset across various metrics in both the All-Search and Indoor-Search scenarios on the different state-of-the-art model such as DDAG Ye et al. [2020b], CAJL Ye et al. [2021a], CMML Ye et al. [2021a], SPOT Chen et al. [2022], DART Yang et al. [2022a], PAENet Zheng et al. [2022], MSCLNet Zhang et al. [2022], DART Yang et al. [2022a], SGIEL Feng et al. [2023], and DEEN Zhang and Wang

Table 7
Performance comparison with different SOTA reID models on SYSU-MM01 dataset

Method	All-Search					Indoor-Search				
	Rank-1	Rank-10	Rank-20	mAP	mINP	Rank-1	Rank-10	Rank-20	mAP	mINP
DDAG Ye et al. [2020b]	54.75	90.39	95.81	53.02	39.62	61.02	94.06	98.41	67.98	62.61
CAJL Ye et al. [2021a]	69.88	95.71	98.46	66.89	53.61	76.26	97.88	99.49	80.37	80.37
CMML Ye et al. [2021a]	69.88	95.71	98.46	66.89	53.61	76.26	97.88	99.49	80.37	76.79
SPOT Chen et al. [2022]	65.34	92.73	97.04	62.25	48.86	69.42	96.22	99.12	74.63	70.48
PAENet Zheng et al. [2022]	74.22	99.03	99.97	73.90		78.04	99.58	100.00	83.54	
MSCLNet Zhang et al. [2022]	76.99	97.93	99.18	71.64		78.49	99.32	99.91	81.17	
DART Yang et al. [2022a]	60.27	93.41	97.47	58.69	45.33	65.74	95.04	98.23	71.77	68.14
SGIEL Feng et al. [2023]	77.12	97.03	99.08	72.33		82.07	97.42	98.87	82.95	
DEEN Zhang and Wang [2023]	74.7	97.6	99.2	71.8		80.3	99.0	99.8	83.3	
CMFT (ours)	82.47	95.66	98.80	75.25	64.67	87.07	97.18	99.54	79.84	69.46

Table 8
Performance comparison with different SOTA reID models on RegDB dataset

Method	Visible to Infrared				Infrared to Visible			
	Rank-1	Rank-10	Rank-20	mAP	Rank-1	Rank-10	Rank-20	mAP
AlignGAN Wang et al. [2019a]	57.90			53.60	56.30			53.40
Xmodal Li et al. [2020]	62.21	-	-	60.18	72.43	-	-	61.80
Hi-CMD Choi et al. [2020]	70.93	-	-	66.04		-	-	-
DDAG Ye et al. [2020b]	69.34	86.19	91.49	63.46	68.06	85.15	90.31	61.80
VI-reID Park et al. [2021]	74.17	-	-	67.64	72.43	-	-	65.46
MPANet Chen et al. [2021b]	83.98	98.54	99.51	81.05	83.69	99.42	99.71	80.72
DAF Ren et al. [2019]	85.39	99.32	99.95	82.11	83.98	99.56	99.95	81.95
CMFT (ours)	89.82	99.30	99.99	85.81	85.41	99.44	99.95	84.28

[2023] in Table7. In the All-Search scenario, the Rank-1 score achieved by our CMFT model stands at 82.47, which is notably higher than any other method. SGIEL [66], the second-best method, has a Rank-1 score of 77.12, indicating an improvement of 7% with our CMFT. The Rank-10 and Rank-20 scores of CMFT are 95.66 and 98.80, which although not the highest, they remain competitive. CMFT's most significant improvement is seen in mAP and mINP scores. Our method achieves an mAP of 75.25, which is approximately 4.1% higher than the next best model DEEN Zhang and Wang [2023] which achieves 71.8. Even more significantly, the mINP score of CMFT is 64.67, well above any other method and shows a remarkable 20.6% improvement over the highest score by CAJL Ye et al. [2021a], which is 53.61.

4.4.3. RegDB

We also assess our approach in comparison to existing competitive methods on RegDB and the focus is on Visible to Infrared and Infrared to Visible performance comparison, as shown in Table 8. In the Visible to Infrared scenario, the CMFT model outperforms all other methods, achieving a Rank-1 accuracy of 89.82%. This is a 4.43% improvement over the second-best method, DAF Ren et al. [2019]. Additionally, the CMFT model attains the highest mAP score (85.81%), representing a 3.7% increase compared to DAF Ren et al. [2019]. For the Infrared to Visible scenario, the CMFT model demonstrates superior performance with a Rank-1 of 85.41%, surpassing the second-best method, DAF, by 1.43%. The CMFT model also exhibits the highest mAP score (84.28%), marking a 2.33% improvement over DAF. Some methods, including AlignGAN Wang et al. [2019a] and Xmodal Li et al. [2020], how considerably lower Rank-1 accuracies and mAP compared to the CMFT model and other SOTA methods.

4.5. Limitations

From the aforementioned analysis, the significance of video-based reID has been established, and our techniques have shown a marked advancement over existing methods. However, our approach requires a pre-determined tracklet length during both the training and testing phases, thus compromising its adaptability in practical scenarios. As future work, we plan to design a pure transformer network capable of processing tracklets of variable lengths.

5. Conclusion

In this study, we introduce a novel transformer-based multimodal person reID fusion model, CMFT, which is specifically designed for person reID tasks that utilize visible-infrared sequences. We leverage the temporal information inherent in video data, allowing for the construction of a comprehensive appearance model for individual identification. The CMFT model is comprised of CCFs, which execute visible-infrared fusion and establish interactions among multi-scales of cross-modality information. Joint training

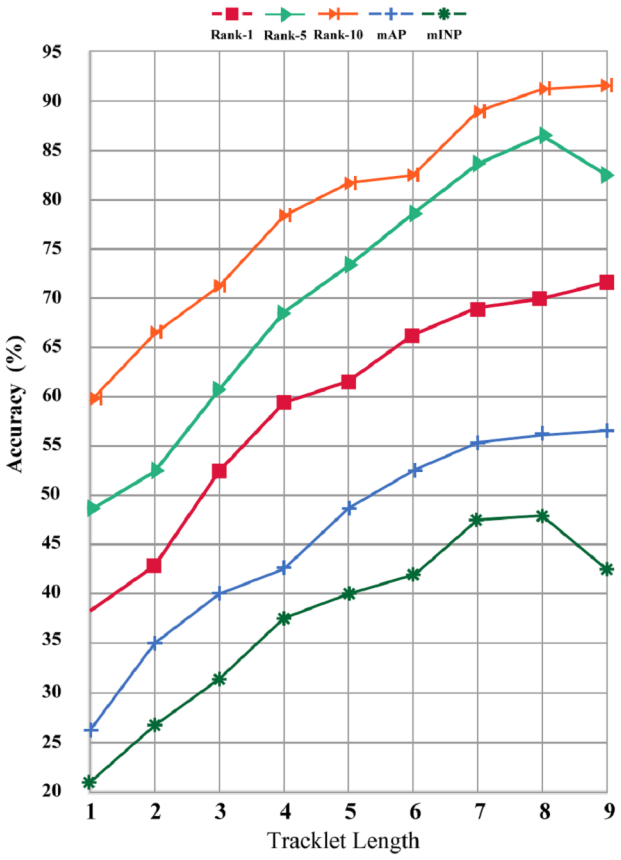

Figure 9: Abalation of CFMT on a different number of tracklet length on HITSZ-VCM dataset.

with dual-stream TSM, the model ensures the exhaustive utilization of complementary information across modalities and efficiently extracts discriminative information. Simultaneously, our proposed MSRL and MDL strategies effectively extract more reliable and distinguishing features, thereby enhancing both the model's performance and robustness. Despite the inherent challenges associated with video-based cross-modal person reID, our proposed technique exhibits remarkable performance. It outperforms existing reID cross-modal methods, demonstrating the effectiveness and potential of our approach in addressing this complex task. The proposed method, by successfully integrating and capitalizing on multimodal and multi-scale information, sets a new standard for in-person reID. In the future, we will focus on investigating more efficient Transformer architectures to further enhance the capacity for feature representation for cross-modality reID.

References

Abhishek Aich, Meng Zheng, Srikrishna Karanam, Terrence Chen, Amit K Roy-Chowdhury, and Ziyan Wu. Spatio-temporal representation factorization for video-based person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 152–162, 2021.

Feng-Ping An and Jun-e Liu. Pedestrian re-identification algorithm based on visual attention-positive sample generation network deep learning model. *Information Fusion*, 86:136–145, 2022.

Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846, 2021.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, volume 2, page 4, 2021.

Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

Cuiqun Chen, Mang Ye, Meibin Qi, Jingjing Wu, Jianguo Jiang, and Chia-Wen Lin. Structure-aware positional transformer for visible-infrared person re-identification. *IEEE Transactions on Image Processing*, 31: 2352–2364, 2022.

Dapeng Chen, Hongsheng Li, Tong Xiao, Shuai Yi, and Xiaogang Wang. Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1169–1178, 2018.

Xianing Chen, Chunlin Xu, Qiong Cao, Jialang Xu, Yujie Zhong, Jiale Xu, Zhengxin Li, Jingya Wang, and Shenghua Gao. Oh-former: Omni-relational high-order transformer for person re-identification. *arXiv preprint arXiv:2109.11159*, 2021a.

Yehansen Chen, Lin Wan, Zhihang Li, Qianyan Jing, and Zongyuan Sun. Neural feature search for rgb-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 587–597, 2021b.

Seokeon Choi, Sumin Lee, Youngeun Kim, Taekyung Kim, and Changick Kim. Hi-cmd: Hierarchical cross-modality disentanglement for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10257–10266, 2020.

Anthony Cioppa, Adrien Deliege, Silvio Giancola, Bernard Ghanem, and Marc Van Droogenbroeck. Scaling up soccernet with multi-view spatial localization and re-identification. *Scientific data*, 9(1):355, 2022.

Srijan Das, Saurav Sharma, Rui Dai, Francois Bremond, and Monique Thonnat. Vpn: Learning video-pose embedding for activities of daily living. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 72–90. Springer, 2020.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Yunhao Du, Zhicheng Zhao, Yang Song, Yanyun Zhao, Fei Su, Tao Gong, and Hongying Meng. Strongsort: Make deepsort great again. *IEEE Transactions on Multimedia*, 2023.

Chanho Eom, Geon Lee, Junghyup Lee, and Bumsub Ham. Video-based person re-identification with spatial and temporal memory networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12036–12045, 2021.

Jiawei Feng, Ancong Wu, and Wei-Shi Zheng. Shape-erased feature learning for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22752–22761, 2023.

Zhanxiang Feng, Jianhuang Lai, and Xiaohua Xie. Learning modality-specific representations for visible-infrared person re-identification. *IEEE Transactions on Image Processing*, 29:579–590, 2019.

Chaoyou Fu, Yibo Hu, Xiang Wu, Hailin Shi, Tao Mei, and Ran He. Cm-nas: Cross-modality neural architecture search for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11823–11832, 2021.

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

Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15013–15022, 2021.

Nianchang Huang, Jianan Liu, Yunqi Miao, Qiang Zhang, and Jungong Han. Deep learning for visible-infrared cross-modality person re-identification: A comprehensive review. *Information Fusion*, 2022.

Kongzhu Jiang, Tianzhu Zhang, Xiang Liu, Bingqiao Qian, Yongdong Zhang, and Feng Wu. Cross-modality transformer for visible-infrared person re-identification. In *European Conference on Computer Vision*, pages 480–496. Springer, 2022.

Shenqi Lai, Zhenhua Chai, and Xiaolin Wei. Transformer meets part model: Adaptive part division for person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4150–4157, 2021.

Diangang Li, Xing Wei, Xiaopeng Hong, and Yihong Gong. Infrared-visible cross-modal person re-identification with an x modality. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 4610–4617, 2020.

Huafeng Li, Minghui Liu, Zhanxuan Hu, Feiping Nie, and Zhengtao Yu. Intermediary-guided bidirectional spatial-temporal aggregation network for video-based visible-infrared person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

Kunchang Li, Yali Wang, Peng Gao, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. Uniformer: Unified transformer for efficient spatiotemporal representation learning. *arXiv preprint arXiv:2201.04676*, 2022.

Yulin Li, Jianfeng He, Tianzhu Zhang, Xiang Liu, Yongdong Zhang, and Feng Wu. Diverse part discovery: Occluded person re-identification with part-aware transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2898–2907, 2021.

Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7083–7093, 2019.

Xinyu Lin, Jinxing Li, Zeyu Ma, Huafeng Li, Shuang Li, Kaixiong Xu, Guangming Lu, and David Zhang. Learning modal-invariant and temporal-memory for video-based visible-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20973–20982, 2022.

Yutian Lin, Lingxi Xie, Yu Wu, Chenggang Yan, and Qi Tian. Unsupervised person re-identification via softened similarity learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3390–3399, 2020.

Deyin Liu, Lin Wu, Richang Hong, Zongyuan Ge, Jialie Shen, Farid Boussaid, and Mohammed Bennamoun. Generative metric learning for adversarially robust open-world person re-identification. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19 (1):1–19, 2023.

Fangyi Liu and Lei Zhang. View confusion feature learning for person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6639–6648, 2019.

Haijun Liu, Xiaoheng Tan, and Xichuan Zhou. Parameter sharing exploration and hetero-center triplet loss for visible-thermal person re-identification. *IEEE Transactions on Multimedia*, 23:4414–4425, 2020.

Yiheng Liu, Zhenxun Yuan, Wengang Zhou, and Houqiang Li. Spatial and temporal mutual promotion for video-based person re-identification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8786–8793, 2019.

Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3202–3211, 2022.

Fuchen Long, Zhaofan Qiu, Yingwei Pan, Ting Yao, Jiebo Luo, and Tao Mei. Stand-alone inter-frame attention in video models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3192–3201, 2022.

Hu Lu, Xuezhang Zou, and Pingping Zhang. Learning progressive modality-shared transformers for effective visible-infrared person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 1835–1843, 2023.

Yan Lu, Yue Wu, Bin Liu, Tianzhu Zhang, Baopu Li, Qi Chu, and Nenghai Yu. Cross-modality person re-identification with shared-specific feature transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13379–13389, 2020.

Tinghuai Ma, Mingming Yang, Huan Rong, Yurong Qian, Yuan Tian, and Najla Al-Nabhan. Dual-path cnn with max gated block for text-based person re-identification. *Image and Vision Computing*, 111:104168, 2021a.

Zhongxing Ma, Yifan Zhao, and Jia Li. Pose-guided inter-and intra-part relational transformer for occluded person re-identification. In *Proceedings of the 29th ACM international conference on multimedia*, pages 1487–1496, 2021b.

Muazzam Maqsood, Sadaf Yasmin, Saira Gillani, Maryam Bukhari, Seungmin Rho, and Sang-Soo Yeo. An efficient deep learning-assisted person re-identification solution for intelligent video surveillance in smart cities. *Frontiers of Computer Science*, 17(4):174329, 2023.

Marina Paolanti, Rocco Pietrini, Adriano Mancini, Emanuele Frontoni, and Primo Zingaretti. Deep understanding of shopper behaviours and interactions using rgb-d vision. *Machine Vision and Applications*, 31: 1–21, 2020.

Hyunjong Park, Sanghoon Lee, Junghyup Lee, and Bumsub Ham. Learning by aligning: Visible-infrared person re-identification using cross-modal correspondences. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12046–12055, 2021.

Aida Pločo, Andrea Macarulla Rodriguez, and Zeno Geradts. Spatial-temporal omni-scale feature learning for person re-identification. In *2020 8th International Workshop on Biometrics and Forensics (IWBF)*, pages 1–5. IEEE, 2020.

Liangliang Ren, Jiwen Lu, Jianjiang Feng, and Jie Zhou. Uniform and variational deep learning for rgb-d object recognition and person re-identification. *IEEE Transactions on Image Processing*, 28(10):4970–4983, 2019.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

Fangmin Sun, Weilin Zang, Raffaele Gravina, Giancarlo Fortino, and Ye Li. Gait-based identification for elderly users in wearable healthcare systems. *Information fusion*, 53:134–144, 2020.

Linfeng Tang, Hao Zhang, Han Xu, and Jiayi Ma. Rethinking the necessity of image fusion in high-level vision tasks: A practical infrared and visible image fusion network based on progressive semantic injection and scene fidelity. *Information Fusion*, page 101870, 2023.

Yusheng Tao, Jian Zhang, Jiajing Hong, and Yuesheng Zhu. Dreamt: Diversity enlarged mutual teaching for unsupervised domain adaptive person re-identifcation. *IEEE Transactions on Multimedia*, 2022.

Xudong Tian, Zhizhong Zhang, Shaohui Lin, Yanyun Qu, Yuan Xie, and Lizhuang Ma. Farewell to mutual information: Variational distillation for cross-modal person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1522–1531, 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF Conference on*

Computer Vision and Pattern Recognition, pages 7464–7475, 2023.

Dongkai Wang and Shiliang Zhang. Unsupervised person re-identification via multi-label classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10981–10990, 2020.

Enqiang Wang, Qing Yu, Yelin Chen, Wushouer Slamu, and Xukang Luo. Multi-modal knowledge graphs representation learning via multi-headed self-attention. *Information Fusion*, 88:78–85, 2022a.

Guan-An Wang, Tianzhu Zhang, Yang Yang, Jian Cheng, Jianlong Chang, Xu Liang, and Zeng-Guang Hou. Cross-modality paired-images generation for rgb-infrared person re-identification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 12144–12151, 2020.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3623–3632, 2019a.

Xianju Wang, Cuiqun Chen, Yong Zhu, and Shuguang Chen. Feature fusion and center aggregation for visible-infrared person re-identification. *IEEE Access*, 10:30949–30958, 2022b.

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 618–626, 2019b.

Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and real-time tracking with a deep association metric. In *2017 IEEE international conference on image processing (ICIP)*, pages 3645–3649. IEEE, 2017.

Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong, and Jianhuang Lai. Rgb-infrared cross-modality person re-identification. In *Proceedings of the IEEE international conference on computer vision*, pages 5380–5389, 2017.

Qiong Wu, Pingyang Dai, Jie Chen, Chia-Wen Lin, Yongjian Wu, Feiyue Huang, Bineng Zhong, and Rongrong Ji. Discover cross-modality nuances for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4330–4339, 2021.

Qiqi Xiao, Hao Luo, and Chi Zhang. Margin sample mining loss: A deep learning based method for person re-identification. *arXiv preprint arXiv:1710.00478*, 2017.

Shuangjie Xu, Yu Cheng, Kang Gu, Yang Yang, Shiyu Chang, and Pan Zhou. Jointly attentive spatial-temporal pooling networks for video-based person re-identification. In *Proceedings of the IEEE international conference on computer vision*, pages 4733–4742, 2017.

Xing Xu, Jialiang Sun, Zuo Cao, Yin Zhang, Xiaofeng Zhu, and Heng Tao Shen. Tfun: Trilinear fusion network for ternary image-text retrieval. *Information Fusion*, 91:327–337, 2023.

Mouxing Yang, Zhenyu Huang, Peng Hu, Taihao Li, Jiancheng Lv, and Xi Peng. Learning with twin noisy labels for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14308–14317, 2022a.

Yulou Yang, Hao Shen, and Ming Yang. Relation-guided network for image-text retrieval. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 1856–1860. IEEE, 2022b.

Mang Ye, Xiangyuan Lan, Qingming Leng, and Jianbing Shen. Cross-modality person re-identification via modality-aware collaborative ensemble learning. *IEEE Transactions on Image Processing*, 29:9387–9399, 2020a.

Mang Ye, Jianbing Shen, David J. Crandall, Ling Shao, and Jiebo Luo. Dynamic dual-attentive aggregation learning for visible-infrared person re-identification. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*, pages 229–247. Springer, 2020b.

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

Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE transactions on pattern analysis and machine intelligence*, 44(6): 2872–2893, 2021b.

JunHo Yoon, GyuHo Choi, and Chang Choi. Multimedia analysis of robustly optimized multimodal transformer based on vision and language co-learning. *Information Fusion*, page 101922, 2023.

Kaiwei Zeng, Munan Ning, Yaohua Wang, and Yang Guo. Hierarchical clustering with hard-batch triplet loss for person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13657–13665, 2020.

Guowen Zhang, Pingping Zhang, Jinqing Qi, and Huchuan Lu. Hat: Hierarchical aggregation transformers for person re-identification. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 516–525, 2021a.

Hao Zhang, Hebaixu Wang, Xin Tian, and Jiayi Ma. P2sharpen: A progressive pansharpening network with deep spectral transformation. *Information Fusion*, 91:103–122, 2023.

Lei Zhang, Fangyi Liu, and David Zhang. Adversarial view confusion feature learning for person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(4):1490–1502, 2020.

Weifeng Zhang, Jing Yu, Wenhong Zhao, and Chuan Ran. Dmrfnet: deep multimodal reasoning and fusion for visual question answering and explanation generation. *Information Fusion*, 72:70–79, 2021b.

Yiyuan Zhang, Sanyuan Zhao, Yuhao Kang, and Jianbing Shen. Modality synergy complement learning with cascaded aggregation for visible-infrared person re-identification. In *European Conference on Computer Vision*, pages 462–479. Springer, 2022.

Yukang Zhang and Hanzi Wang. Diverse embedding expansion network and low-light cross-modality benchmark for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2153–2162, 2023.

Aihua Zheng, Peng Pan, Hongchao Li, Chenglong Li, Bin Luo, Chang Tan, and Ruoran Jia. Progressive attribute embedding for accurate cross-modality person re-id. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4309–4317, 2022.

Aihua Zheng, Xianpeng Zhu, Zhiqi Ma, Chenglong Li, Jin Tang, and Jixin Ma. Cross-directional consistency network with adaptive layer normalization for multi-spectral vehicle re-identification and a high-quality benchmark. *Information Fusion*, 100:101901, 2023.

Zhedong Zheng, Liang Zheng, and Yi Yang. A discriminatively learned cnn embedding for person reidentification. *ACM transactions on multimedia computing, communications, and applications (TOMM)*, 14(1):1–20, 2017.

Kuan Zhu, Haiyun Guo, Shiliang Zhang, Yaowei Wang, Gaopan Huang, Honglin Qiao, Jing Liu, Jinqiao Wang, and Ming Tang. Aaformer: Auto-aligned transformer for person re-identification. *arXiv preprint arXiv:2104.00921*, 2021a.

Xingkui Zhu, Shuchang Lyu, Xu Wang, and Qi Zhao. Tph-yolov5: Improved yolov5 based on transformer prediction head for object detection on drone-captured scenarios. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2778–2788, 2021b.