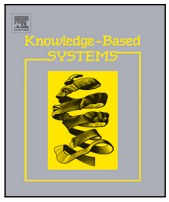
Knowledge-Based Systems 283 (2024) 111155

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys



CMOT: A cross-modality transformer for RGB-D fusion in person re-identification with online learning capabilities

Hamza Mukhtar *, Muhammad Usman Ghani Khan

Department of Computer Science, University of Engineering and Technology Lahore, G.T. Road, Lahore, 54890, Punjab, Pakistan
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:
Person re-identification
Fusion transformer
Cross-modality
Few-shot learning
RGB-D sensors

ABSTRACT

Person re-identification (reID) is a crucial aspect of intelligent surveillance systems, enabling the recognition of individuals across non-overlapping camera views. As compared to the RGB modality, RGB-D based reID has the potential to achieve robust and high performance by leveraging rich complementary features of both modalities, making it applicable in various occluded scenarios. However, current multimodal reID approaches often rely on late fusion or feature-level fusion techniques to combine multiple modalities, which limits their ability to proficiently exploit complementary visual and depth-related semantic information and capture complex interactions between unimodal features. To address these limitations, this paper introduces a cross-modality online transformer (CMOT) for RGB-D based person reID with online learning capabilities, which effectively utilizes both RGB and depth modalities for the extraction of spatio-temporal features and fuses across modalities. Our CMOT is composed of three main components: (1) a hypothesis generation module based on a person detector and tracker, (2) dual-stream feature extractors via convolutional neural networks (CNNs), (3) and a fusion transformer based on a self-attention-driven self-attentive modality refinement module (SAMR) and a cross-attention-driven cross-attentive modality interaction module (CAMI) to refine and fuse RGB-D complementary features extracted from the dual-stream, RGB and depth stream, CNNs. Additionally, we introduce a bottleneck enhancement feed-forward block to enhance the model's representation capability within SAMR and CAMI, significantly reducing parameters and computations compared to the traditional feed-forward network. Moreover, we design the triplet loss function with distance measuring ability for incorporating online learning and finally CMOT works as a few-shot network for reID. Experimental results on three RGB-D person re-identification datasets, namely BIWI RGBD-ID, RobotPKU RGBD-ID, and TVPR2, demonstrate the effectiveness and robustness of CMOT.

1. Introduction

Person re-identification (reID) [1–3], the process of identifying individuals across distinct, non-overlapping camera views, is critical for intelligent video surveillance systems, such as assisting in forensic investigations, facilitating multi-camera tracking, improving access control measures, and contributing to sports analytics. Recent applications of reID have included service robots and human–robot interactions, particularly for monitoring and aiding the elderly with individualized activities [4]. Despite its broad application, variables like changing viewing angles, lighting, stance discrepancies, occlusions, and background clutter all impede consistent recognition across different camera perspectives [5]. The development of precise and cost-effective depth sensors, such as Kinect [6], Asus's Xtion Pro Live [7], and Intel RealSense [8] has led to a surge of interest in RGB-D-based reID across various application domains, including intelligent video surveillance [1], human action recognition [2], and shopper behaviour analysis in physical stores [3]. Mirroring human sensory perception, current studies [9,10] highlight the value of multi-modal information in classification tasks and explore interconnections between different modalities, which significantly improves performance. Consequently, the diverse and heterogeneous modalities present in video data can be harnessed to provide a more comprehensive feature representation, particularly in the context of person reID. For example, the RGB and depth modalities provide complementary features where the depth modality is rich in 3D structural data and remains unaffected by variations in lighting. However, it falls short when it comes to providing essential appearance information. On the other hand, the RGB, while exhibiting the opposite characteristics [11], provides valuable colour and texture

* Corresponding author at: Department of Computer Science, University of Engineering and Technology Lahore, G.T. Road, Lahore, 54890, Punjab, Pakistan.
E-mail addresses: hamza.mukhtar@kics.edu.pk (H. Mukhtar), Usman.ghani@uet.edu.pk (M.U.G. Khan).

https://doi.org/10.1016/j.knosys.2023.111155
Received 23 August 2023; Received in revised form 23 October 2023; Accepted 30 October 2023
Available online 2 November 2023

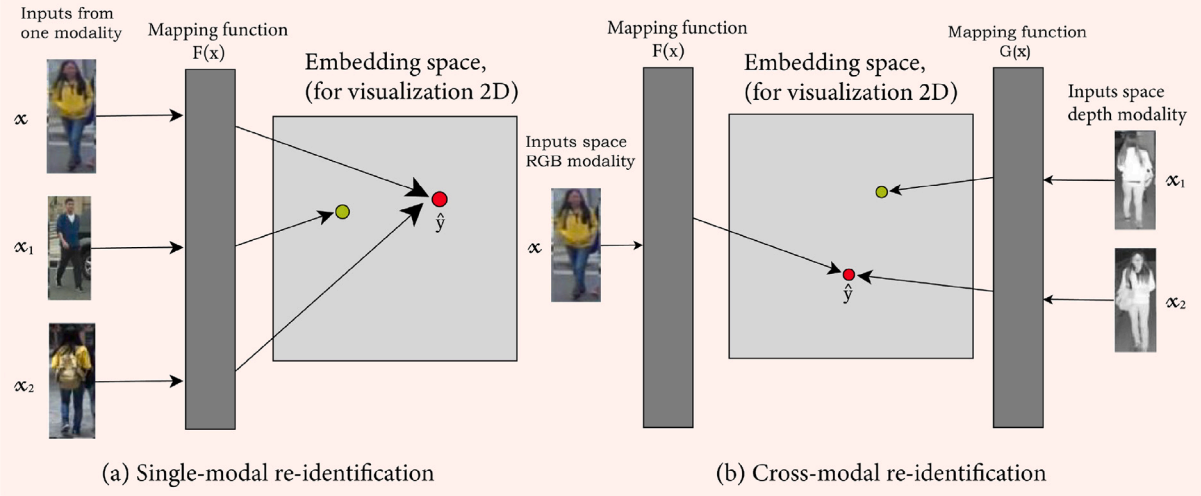


Fig. 1. (a) In single-modal reID, an embedding function F(x) maps identical modal inputs into a common feature space, grouping same-person images closely. x is the query image paired with gallery images x_1, x_2 (one sharing x's class), with $\hat{y}$ as the nearest embedding to x. (b) Cross-modal re-ID uses individual mapping functions (F(x) for RGB, G(x) for depth) to generate shared embeddings across modalities. The red dot refers to matched embedding between query and gallery inputs, while the green dot shows unmatched with query input.

details that the depth modality lacks. Despite these complementary attributes, previous research [12–14] has predominantly focused on employing a single modality for reID, overlooking the potential benefits of exploiting the complementary characteristics of different modalities.

Most existing person reID techniques concentrate on matching persons and can be categorized as appearance-modality based [12–14] or multi-modality-based [15,16]. Methods based on appearance-modality [17–21] primarily focus on robust feature descriptors that can adapt to changes in lighting, pose, and viewing angle. They also aim to establish discriminative similarity metrics for person matching. However, these methods may encounter difficulties in certain scenarios due to substantial intra-class divergence (such as individuals viewed from different angles or under varying lighting conditions) and minimal inter-class divergence (such as individuals wearing similar clothing). This is particularly evident in environments like schools where uniforms are worn. In contrast, multi-modality-based approaches [22,23] emphasize the integration of appearance information with other modalities. These include thermal data [24], gait analysis [25], and anthropometric measurements of the body [24,26,27], which are resilient to changes in lighting conditions, viewpoints, and clothing.

However, existing approaches in RGB-D person reID often rely on a single trained model that processes both RGB and depth modalities, with depth images converted into 3-channel representations, in order to enhance reID performance. A dual-stream network [28] processes the RGB and RGB-depth input with separate CNN to generate distinct feature embeddings which are necessary for pairwise matching with reference image embeddings. The features extracted from the final fully connected layer of both CNNs are then fused to facilitate joint reID learning. While these multi-modal models have shown promising outcomes on publicly available datasets, their performance might prove inadequate when compared with significantly varying environments. This limitation can be attributed to two primary reasons. Firstly, many conventional approaches primarily train similarity metric models offline, thereby lacking the capacity to adapt to novel scenes that deviate substantially from the training data distribution due to changes in illumination conditions, camera viewpoints, or backgrounds. Secondly, the naive combination of multiple features for calculating person similarity can lead to error accumulation and introduce unnecessary computational overhead. Fig. 1 illustrates the distinction between single-modality (RGB) and multi-modality (RGB-D) person reID approaches.

Vision-transformer [29] has gained attraction in computer vision tasks due to its capacity to capture long-range dependencies and less restrictive inductive biases [30]. In reID, transformers have primarily been employed to focus on capturing long-term information and spatio-temporal context [30–33]. However, the application of transformers to multimodal data, which is common in visual-linguistic tasks and autonomous driving, has received limited attention. While there have been some attempts to utilize transformer-based components for RGB-D reID [34,35], the fusion module employed in these approaches falls short of effectively modelling inter-modal relationships. As a result, the full potential of transformers in the context of RGB-D-based multimodal fusion has not been fully harnessed.

Based on the above observations, we propose CMOT, a reID model for RGB-D that consists of a hypothesis generation module, a dual-stream feature extraction CNN, and a fusion transformer to fuse the cross-modality information. The hypothesis generation module comprises two main components: a person detection model, YOLOv5 [36], and a tracking algorithm, StrongSort [37]. Together, these components contribute to generating a sequence of person-specific regions, effectively identifying and tracking person instances within the RGB and depth sequence. CMOT aims to jointly learn spatio-temporal information and execute feature-level multimodal fusion in an integrated framework. By leveraging modality-specific and cross-modal interactions, CMOT enhances the discriminative ability of the learned representations for accurate and robust reID in RGB-D scenarios. The fusion transformer in CMOT is constructed with a self-attentive Modality Refinement Module (SAMR) and a cross-attention-based Cross-Attentive Modality Interaction Module (CAMI), similar to work [38]. SAMR facilitates the adaptive enrichment of semantic information within each stream's features, while CAMI processes the features from both streams to acquire cross-modal complementary features. Consequently, CMOT alternately fuses and enhances multimodal information by leveraging the capabilities of self-attention and cross-attention. In order to enhance the expressive capacity of CMOT, we propose the incorporation of a bottleneck enhancement feed-forward block (BEFFB) as a substitute for the conventional feed-forward network (FFN) employed in the vanilla transformer [39]. The BEFFB module is specifically designed to reduce the parameter count while preserving or even enhancing performance, thereby addressing the challenge of achieving a balance between model complexity and computational efficiency. Furthermore, we exploit the rank pooling mechanism, as introduced in prior works [40,41], to depict RGB-D sequences as unimodal visual dynamic sequences (VDSs) and depth dynamic sequences (DDSs). This approach facilitates the conversion of temporal information into spatial representations encapsulated in the sequences, thereby empowering the employment of CNNs

for the extraction of features. By converting the sequences into VDSs and DDSs, we can leverage the spatial-awareness capabilities of CNNs to capture modality-specific discriminative visual patterns and depth-based cues that encode appearance and geometric cues from the RGB and depth data. Traditional one-and-done person reID approaches [32,34,35,39,42,43] face challenges in contexts requiring incremental data acquisition, with standard paradigms necessitating retraining of reID models, leading to inefficiency and practical storage issues [42]. This motivates us to focus on reID tasks that efficiently utilize new data acquired in a piecemeal fashion. However, incremental training risks catastrophic forgetting [44,45], reducing discriminatory capabilities on unseen data. To mitigate this, we formulate a similarity-based triplet loss strategy, transitioning our reID system to an online learner model through a few-shot learning network via distance learning.

This study makes several contributions to the field of multi-modality reID:

1. We introduce a transformer-based multi-modality reID model, CMOT, that uses both RGB and depth modality to concurrently execute spatiotemporal information extraction and cross-modality fusion, comprising dual-stream CNNs and a fusion transformer. This framework allows for the utilization of complementary and discriminative features extracted from multi-modal dynamic sequence input through dual-stream CNN.
2. We develop a self-attention-driven SAMR and a cross-attention-driven CAMI to facilitate cross-modal feature refinement and interaction, respectively. The SAMR enriches the semantic information within each modality, while the CAMI processes features from both streams to learn cross-modal complementary features. By combining SAMR and CAMI as a single basic fusion layer within CMOT, we achieve efficient feature-level multi-modal fusion. We also present the BEFFB to improve the model's representation capability, considerably decreasing the model parameters and eventually computational operations as compared to the FFN.
3. We propose a similarity distance measure based triplet loss function and transform CMOT into a few-shot network for incorporating online learning for reID inference. This enables incremental updates and adaptation to new unseen data without model retraining.
4. Evaluation on four datasets, such as BIWI RGBD-ID [46], RobotPKU RGBD-ID [47] and TVPR2 [3], with varying viewpoints and complexity shows that CMOT is effective in capturing the complementary features from cross-modalities.

The remaining article is organized as follows: Section 2 reviews the different types of previous reID methods, such as Appearance-modality reID, multi-modality-based reID, vision-transformer-based reID, and online learner reID, and identifies the research gap. Section 3 presents our proposed CMOT, 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

This section discusses various methods that are classified into four broader categories, including Appearance-modality-based reID, Multi-modality-based reID, transformer-based reID, and online learning mechanism in reID.

2.1. Appearance-modality-based reID

Most existing reID [12,13,48,49] methods concentrate on matching individuals using conventional RGB cameras. Traditional reID methods focus on single-modality reID using RGB-RGB images to match individuals across different cameras with no overlapping views. The reID task poses significant challenges in learning discriminative features from the person images, primarily attributed to the substantial intra-class variations and limited inter-class variations. These variations arise from various factors such as diverse poses, varying illumination conditions, and occlusions resulting from different viewpoints. In response to these challenges, numerous reID methods [15,50] have been proposed. These reID methods encompass various techniques, including the utilization of loss functions, which play a pivotal role in feature learning. Specifically, contrastive [17], triplet [51], and quadruplet loss [52] are among the popular loss functions employed in deep reID methodologies. These loss functions aim to optimize the feature representations by encouraging enhanced discrimination between images of different individuals (inter-class variations) while promoting compact clusters for images of the same individual (intra-class variations).

Graph-based methods for reID [53,54] incorporate the relationship between pairs of samples to enhance the matching accuracy. These methods leverage graph structures to model the connections between samples and capture the underlying associations between individuals. On the other hand, some reID approaches utilize the aligned body parts to improve the discriminatory capabilities of the learned features [15,18]. PDGCN [55] uses a priority-based graph convolution network (PGCN) and a priority distance (PD) where each feature vector serves as a graph node, with neighbouring nodes aggregated to derive context-aware, globally embedded features. PGCN predicts node priority as a class centre by leveraging contextual information, while PD computes distances between high-priority nodes to identify reliable clustering centres. The PCB [18] employs a uniform partition mechanism to extract features at the part level, capturing both global and local information. Refined part pooling enhances within-part consistency and assigns outliers to their corresponding semantic parts. The MGCA [16] captures representations with varying granularity by utilizing global and local branches. It incorporates multi-scale features to enhance the discriminative power. The MPN [19] leverages coarse prior information about body part locations, guiding feature extraction with a mask-guided mechanism. Main and auxiliary tasks improve feature quality and capture fine-grained details. These techniques emphasize the importance of granularity and contextual information in reID. However, part-based methods face challenges in cross-modality scenarios [1,48]. The challenges faced in cross-modal settings for part-based approaches in reID arise from variations in appearance, imaging conditions, and data distributions across different modalities. These factors can potentially impact the performance of such methods. Moreover, existing reID techniques often fail to fully utilize the abundant multimodal information present in video data. In contrast, our CMOT leverages the integration of multiple modalities to enhance robustness and accuracy in reID tasks.

2.2. Multi-modality-based reID

Studies have proposed integrating appearance information with other modalities such as thermal, RGB-D [20], gait information [21], and anthropometric features [34,41] to enhance the accuracy of appearance-based reID techniques. Anthropometric measures encompass various aspects of the human body, such as size, shape, length, width, and height. In multimodal approaches [22], skeleton-based features derived from anthropometric measurements are utilized. By incorporating these features, multimodal systems can capture the structural characteristics of the human body for improved person reID performance. Moreover, a multi-modality system [24] has been proposed that integrates RGB, depth, and thermal information. The modalities are merged using a late fusion strategy, where the information from different modalities is combined at a later stage. In RGB-D modality, SeSAME [25] has demonstrated that the inclusion of anthropometric measurements, combined with different clothing appearance descriptors, can effectively enhance reID. These findings emphasize the prospective integration of anthropometric data to augment the

discriminative capabilities of reID systems, particularly when coupled with appearance-based features.

Despite the simultaneous capturing of RGB-D and skeleton features by RGB-D sensors, challenges may arise in scenarios characterized by significant variations in lighting conditions or clothing, necessitating specialized approaches for RGB-D-based reID [32,39,42,56] to address these constraints. Recurrent Attention Model (RAM) [56] is designed specifically for reID using depth modality. This approach employs 3D RAM for static point clouds and extends it to 4D RAM for 3D point cloud sequences. However, it is important to highlight that this approach might not be the best fit for situations where there is a lack of identity overlap between the training and testing datasets. Another alternative [39] introduces a long-term reID method that utilizes depth videos. Temporal information is leveraged in [57] through a reinforced temporal attention mechanism applied to frame-level features. This mechanism adaptively weights frame predictions based on a task-based reward, enhancing the utilization of temporal information. Feature extraction from depth images and skeleton joint points is performed in [33,56], with score-level fusion techniques used to enhance reID. These depth-based approaches address the challenges posed by RGB, offering potential solutions for robust reID in dynamic scenarios. In depth-based reID [41], a depth voxel covariance descriptor and a rotation-invariant eigen-depth feature are developed, complemented by skeleton-based features. Similarity measurement is performed by combining Euclidean and Geodesic distances. Moreover, in [41], depth images are utilized along with anthropometric features. Local binary, derivative, and tetra patterns are extracted from depth images, which are segmented into head, torso, and legs using skeleton data and fused with anthropometric features.

Multimodality systems for reID can be categorized into two groups [58]. The first category involves fusing information at the feature level [59,60]. This approach concatenates feature vectors from different modalities to form a final feature representation. However, this method often neglects the varying reliability and importance of different features, potentially leading to suboptimal fusion. The second category focuses on fusing information at the score level [11, 47]. This approach combines the scores or predictions from different sub-systems or modalities. However, previous methods based on score-level fusion have certain limitations. The primary objective of these approaches is to minimize the dissimilarity between modalities by learning representations that exhibit greater similarity across different modalities. One specific approach [10] employs the design of single-stream networks, such as modality fusion learning networks. These networks introduce domain-specific nodes to facilitate the alignment of feature representations across different modalities. To learn multi-modality representations in reID, frameworks have developed dual-stream networks that leverage modality-specific architectures [61] and top-ranking loss functions [62]. These approaches aim to effectively capture and utilize information from different modalities for improved reID. In addition, a GAN is used in [63] to jointly discriminate identity and modality, enabling the learning of discriminating representations that account for both appearance and modality information. Furthermore, Xiang et al. [63] introduced a hetero-centre loss that focuses on minimizing intra-class variations by enforcing a reduced distance between intra-class samples of different modalities. This loss encourages the model to learn representations that are robust to variations across modalities.

Cross-modal reID models have gained significant focus due to their practicality in real-world scenarios characterized by varying illumination conditions. Various works are exploring GANs to create depth to RGB modality and RGB to depth modality [64], and reID specific loss functions [63] which are directed towards the encoding of RGB and IR modality images into a shared feature space. TSLFN [63] integrates modality-specific information and optimizes it through an auxiliary classifier. AXM-Net [65] combines the textual and image modalities, where textual input specifies the description of the person related to the corresponding image. In recent studies, deep learning techniques have been experimented to tackle RGB-D reID [28,49,66]. These methods employ feature-level fusion strategies to integrate multi-modal features from RGB-D sensors. One such approach [49] employs a multi-modal fusion layer to merge depth and RGB appearance information, while another study [66] proposes a deep network with a cross-modality auto-encoder incorporated at the top of the network. Deep learning techniques and feature-level fusion strategies in RGB-D reID demonstrate the ongoing exploration of leveraging multi-modal data to enhance reID performance.

Recent works have given rise to cross-modality datasets employing RGB-D sensors and infrared cameras [10,26,27] to evaluate multi-modal methods. Cross-modal models aim to achieve shared common feature representations by employing either single-stream [47,59] or dual-stream networks [16,34]. These models try to bridge the inherent gaps between different modalities. Additionally, some techniques generate an intermediate modality [24,25] to counteract the effects of modality discrepancy. A similar strategy employs GAN techniques [19] to create cross-modality images for person matching. However, generating common modality images adds noise due to imperfect mapping and data distribution differences. Furthermore, the generated images substantially increase computational demands and add more uncertainty to cross-modality learning.

2.3. Transformer-based reID

Humans process data selectively, focusing on salient parts of information to make decisions [67,68]. This process, commonly referred to as the attention mechanism, has found extensive application in various tasks, including image captioning [69] and object detection [70]. In the field of computer vision, the self-attention mechanism [71] is introduced to capture global dependencies within input data. Several methods [70,72] have leveraged self-attention to improve the classification localization models. In recent years, the application of transformers in reID has gained attention, particularly in supervised settings. In reID systems, attention mechanisms are commonly employed to capture spatial and temporal features across video frames [72]. Early approaches [12,48,73] frequently combined transformer layers with CNN feature extractors to effectively capture fine-grained cues and long-range contextual information. For instance, the TPM method [12] introduced a transformer-based module to adaptively merge parts extracted from CNN-based networks such as PCB [15] and MGN [74]. Another method, HAT [73], developed a hierarchical aggregation transformer on top of ResNet-50 [75] to integrate low-level details with high-level semantics. PAT [76] incorporated a part-aware transformer following a CNN backbone to identify diverse parts within-person images. More recently, pure transformer architectures designed specifically for supervised reID tasks have also emerged, where TransReID [77], AAformer [43], TransVI [78] and Denseformer [79] are notable examples of such architectures. AAformer [43] introduces part tokens within the transformer to learn part features, enhancing the model's ability to capture part-level details, and TransVI [78] processed the RGB-Infrared modalities to explicitly capture the modality-specific representations and learn multi-modality sharable knowledge. On the other hand, Denseforme [79] employs a strategy similar to PCB for extracting part-level features. These advancements in transformer-based architectures demonstrate the exploration and integration of attention mechanisms in person reID. By combining transformers with CNN feature extractors or adopting pure transformer architectures, researchers aim to learn fine-grained features, long-range dependencies, and part-level information for improved supervised reID performance.

In cross-modality reID, various methods have been proposed to capture invariant information and establish a shared feature space across different modalities. For instance, a multi-granularity attention network [43] to achieve this goal. In contrast, HRN [14] adopts a different approach by utilizing a dual-modal branch that specifically

targets the capture of shared spatial contexts between modalities. This branch incorporates shared attentive pooling and mutual contextual graph networks, which collectively enable the extraction of spatial attention within individual local areas as well as the modelling of spatial relationships between distinct local components. By incorporating these mechanisms, HRN aims to enhance the representation learning process by effectively leveraging spatial information across different modalities. However, despite the successes of these methods, they still encounter challenges in cross-modality reID, particularly in scenarios with dramatic visual appearance changes across different camera environments. To address these limitations, our proposed approach shares similarities with existing RGB-D cross-modal pedestrian re-identification methods [34,35,69] in terms of adopting a two-stream network and aiming to capture cross-modal relationships. However, a key distinction lies in the approach to extracting cross-modal relationships. Our method introduces the use of external cues for feature partitioning, allowing the model to simultaneously perceive coarse and fine features through distinct self-attention and cross-attention mechanisms. These cues aid in feature partitioning, enabling the model to perceive both coarse and fine features through self-attention and cross-attention mechanisms. They facilitate capturing dependencies between RGB and depth modalities, improving cross-modal reID performance significantly. A transformer network for the RGB and depth modality is used to fuse these modalities and extract external cues through the fusion of RGB and depth through self-attention and cross-attention mechanisms. This approach enables the model to effectively capture the dependencies between modalities and leverage both global and local information for enhanced cross-modal reID performance. By utilizing external cues and incorporating specialized attention mechanisms, our method expands upon the existing approaches in order to tackle the challenges associated with cross-modal reID.

2.4. Online learning

Most traditional reID methods typically learn the similarity metric model offline which poses challenges in adapting to unknown scenes [14]. To address this limitation, semi-supervised learning methods have gained popularity in various computer vision applications, including person reID [80,81]. Various works are exploring various aspects of this field, including network architecture [61], and distance learning to measure the similarity between support set and query input [62] These methods leverage both labelled and unlabelled data in the training process, allowing for enhanced performance by incorporating information from unlabelled data. In particular, distance learning [82] has proven to be an effective approach for online learning in reID. It utilizes positive and negative constraints to guide the labelling process for unlabelled data, enabling the model to learn and refine its representations based on the constraints imposed by these labelled samples. In this paper, we propose incorporating distance learning into our online reID framework. By incorporating distance learning techniques, we aim to measure the similarities between the query image and support images, enabling us to predict the most relevant matches based on learned distance metrics. By leveraging both labelled and unlabelled data, our approach facilitates online learning and enhances the adaptability of the reID system to handle unknown scenes.

3. Methodology

For effective extraction and integration of RGB and depth modality features for reID, we present the CMOT, as illustrated in Fig. 2. The proposed online learning mechanism-based cross-modality re-identification framework consists of four components: (1) Hypothesis generation module that detects and trackers person instances to form a sequence of each person from RGB and depth input modality, (2) dual-stream CNNs to process the person's RGB-depth sequence separately and extract the features fine-grained and coarse-grained features, (3) cross-attention fusion transformer for cross-modality information fusion to learn self and cross-modality dependencies, and (4) output layer that receives the fused features of both modalities and uses cosine similarity to measure the similarity between query and support sequence.

The fusion of RGB and depth information enhances the system's robustness, particularly in challenging scenarios. Occluded scenarios, where individuals may be partially hidden or obstructed by objects or other people, pose a significant challenge to person re-identification systems. In such situations, depth information becomes invaluable. It can distinguish between overlapping individuals by distinguishing the spatial relationships between body parts. For instance, it enables the recognition of a person even when only a portion of their body is visible, contributing to performance enhancement and robustness.

3.1. Hypothesis generation

In a video, there are multiple person instances and background regions as well. To generate the region-of-interest (ROI) hypothesis comprised of only the person's body region, our approach incorporates YOLOv5 [36], an end-to-end detection model, in combination with StrongSORT [37] for tracking purposes. YOLOv5 is composed of three main components: the backbone for feature extraction, the neck network for processing the extracted features, and the head layer to output the bounding box and confidence. The backbone uses a CNN to capture the low- and high-level geometrical and visual features. The neck component processes features extracted from the backbone CNN to generate feature maps that are subsequently used for prediction. The YOLOv5 architecture employs CSPDarknet53 as its backbone network, which comprises 29 convolutional layers with a 3×3 kernel size, resulting in a receptive field of 320×320 pixels and a total of 27.6 million parameters. To increase the receptive field without sacrificing computational efficiency, a Spatial Pyramid Pooling (SPP) block is incorporated into CSPDarknet53. Additionally, the Path Aggregation Network (PANet) is employed to fuse low- and high-level features, thereby enhancing the richness and discriminative power of the extracted features. This integration of YOLOv5 with StrongSORT facilitates accurate object detection and tracking, enabling the generation of reliable region-of-interest hypotheses for further processing.

For the tracking stage, we utilize StrongSORT [37], an enhanced version of the DeepSORT algorithm [83]. DeepSORT builds upon the SORT algorithm [84], which performs object tracking by utilizing the Kalman Filter [42] for motion prediction and the Hungarian algorithm for identity association across consecutive frames. However, SORT encounters difficulties when objects are occluded, resulting in incorrect associations and identity switches. To address these challenges, DeepSORT incorporates a pre-trained CNN network to capture the appearance features of objects over the last 100 frames, mitigating identity switches caused by occlusions. Additionally, DeepSORT introduces cascade matching and trajectory confirmation mechanisms to enhance trajectory prediction and object matching in the current frame. StrongSORT, an extension of DeepSORT, incorporates two additional lightweight algorithms: AFLink and GSI. AFLink associates short trajectories with complete trajectories using a fully connected model without relying on appearance information. This mechanism improves trajectory completeness and reduces fragmentation. GSI enhances detection reliability by simulating nonlinear motion patterns and leveraging Gaussian regression to achieve more accurate object positioning. Importantly, GSI considers the motion information of detected objects during the regression process, enabling more robust and accurate trajectory estimation.

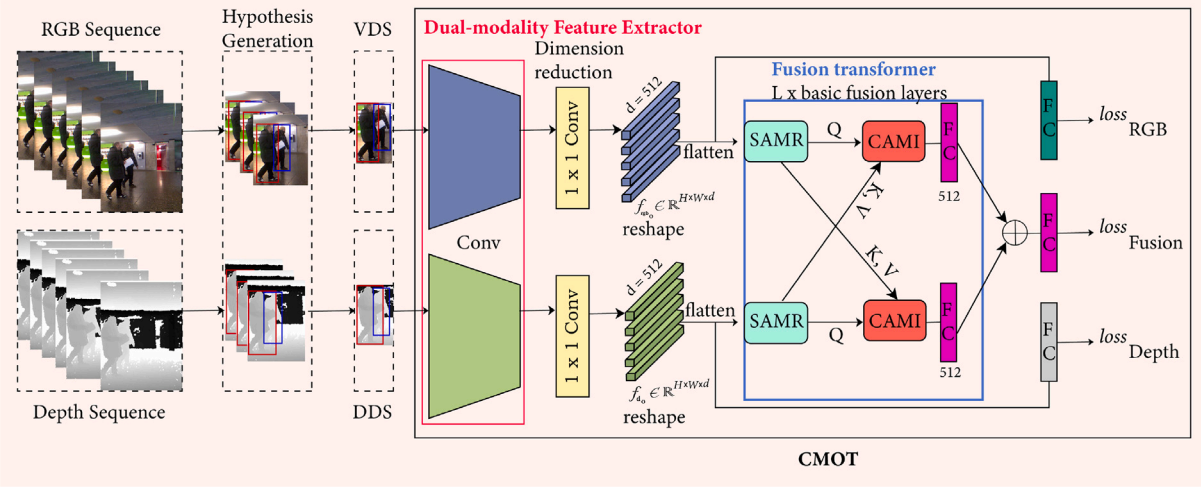


Fig. 2. The proposed CMOT has three core modules: (1) a hypothesis generation module that detects and tracks the person instances and produces the pair of dynamic sequences $(VDS, DDS) \in R^{3\times H\times W}$ from RGB and depth frame sequence of each person, (2) a dual-stream feature extractor that receives dynamic sequences (VDS, DDS) as input and extracts the unimodal features $(f_{rgb}, f_d) \in R^{C\times H\times W}$, (3)and a fusion transformer that takes unimodal features as input $(f_{rgb1}, f_{d1}) \in R^{HW\times d}$ for cross-modality feature fusion. Finally, the feature vectors from the RGB, Depth, and RGB-D branches are fed to the FC layer. Fusion of both RGB and depth sequences offers the key advantage of the RGB-D modality in handling occluded pedestrian reID. The depth sequence, which provides depth information alongside appearance, plays a pivotal role in enhancing the system's performance in scenarios with occluded pedestrians. By leveraging the depth information, the system can better distinguish the spatial relationships between objects, allowing it to detect pedestrians even when partially occluded. This inherent capability of the RGB-D sequence is a fundamental factor contributing to the robustness and high performance of our approach in occluded scenarios.

3.2. Dual-modality feature extraction

In contrast to traditional unimodal convolution networks [10,61,64] that extract features from each modality separately, the proposed CMOT employs two distinct CNNs to extract fine-grained and coarse-grained features from RGB and Depth modalities simultaneously. Dynamic images, which effectively represent spatio-temporal information for person reID, have been validated in previous studies [3,34,66]. As a result, we use these representations, created by encapsulating the hypothesis frame sequence into a dynamic image, as the input to our feature extractor using the temporal rank pooling mechanism. In Fig. 2, a pair of dynamic RGB and depth sequence $(VDS, DDS) \in R^{3\times H\times W}$ derived from the hypothesis sequence, is input into the dual-modality feature extractor. RGB and depth modality feature extractors focus on extracting spatiotemporal information from RGB and depth dynamic sequences, respectively. Given the impressive achievements of the ResNet architecture [75] across diverse visual tasks, we incorporate ResNet50 for each stream backbone, facilitating the extraction of spatio-temporal features. We have removed the final stage of ResNet50 for the effective integration of the cross-modality features because this stage is predominantly designed to handle unimodal inputs and may not effectively fuse the features from our multi-modality inputs. Instead, we utilize the output from the fourth bottleneck as the terminal output for our dual-modality feature extractor. This adjustment aligns better with our aim of integrating, harnessing, and integrating the strengths of different modalities. Our dual-modality feature extractor, with these modifications, processes the RGB and depth dynamic sequences simultaneously. From each of these modalities, it produces corresponding feature maps $(f_{rgb}, f_d) \in R^{C\times H\times W}$, rich in spatio-temporal information.

3.3. Fusion transformer

Current approaches typically rely on fusing prediction scores from cross-modalities to form the final outputs [14,35] or simply concatenate the flatten feature maps obtained from the last convolutional layer of the feature extractor, CNN, to achieve a basic level of cross-modality fusion [31,60]. However, these strategies may not fully exploit the synergistic potential offered by the combination of different modalities. Inspired by the success of cross-modality fusion capabilities of cross-attention used in visual question-answering [85,86] and human action recognition [2], we propose an advanced fusion strategy — the cross-modality transformer. This approach is specifically designed to harness and capitalize on the complementary aspects of visual and depth modalities. As depicted in Fig. 2, our fusion transformer incorporates two primary modules: the modality refinement module, SAMR, and the interaction module, CAMI. The SAMR module focuses on strengthening the individual characteristics and features of each modality. Concurrently, the CAMI module encourages intercommunication between the different modalities, enabling the model to detect and utilize the complementary and distinctive features of the RGB and depth data. Through this approach, the fusion transformer optimizes the fusion of the different modalities and facilitates the extraction of more meaningful and discriminative features for reID. RGB images provide rich texture and colour information, while depth maps offer valuable insights into the spatial structure and three-dimensional shape of the subjects. The highly complementary nature of both modalities offers a more comprehensive representation of the person of interest. However, establishing connections between these two inherently heterogeneous types of features poses a significant challenge. The information both modalities provide resides in distinctly different domains — RGB images in the colour and texture domain and depth maps in the 3D spatial structure domain. This difference introduces a level of complexity when attempting to integrate these disparate features in a meaningful way.

The cross-modality fusion process is initiated with the application of 1×1 convolution on the input feature maps $f_{rgb} \in R^{C\times H\times W}$ and $f_d \in R^{C\times H\times W}$, which are obtained from the RGB and depth dynamic images, respectively. This convolution operation is intended to decrease the channel dimensions of the input feature maps. As a result, it generates new feature maps $f_{rgb0} \in R^{H\times W\times d}$ and $f_{rgb1} \in R^{H\times W\times d}$, where the channel dimension is lower than that of the original inputs. Following this, a flattening operation is applied to these transformed feature maps along their spatial dimensions, converting them into one-dimensional spatial features. This operation yields the transformed inputs $f_{rgb1} \in R^{HW\times W}$ and $f_{d1} \in R^{HW\times W}$ for the CMOT.

Dimension d represents the feature embedding size, and the one-dimensional spatial feature size, which corresponds to the sequence length, is pivotal in maintaining the spatial–temporal context of the input sequence. These unimodal inputs are passed sequentially into the Self-Attentive Modality Refinement Module (SAMR) and the Cross-Attentive Modality Interaction Module (CAMI). SAMRs, employing

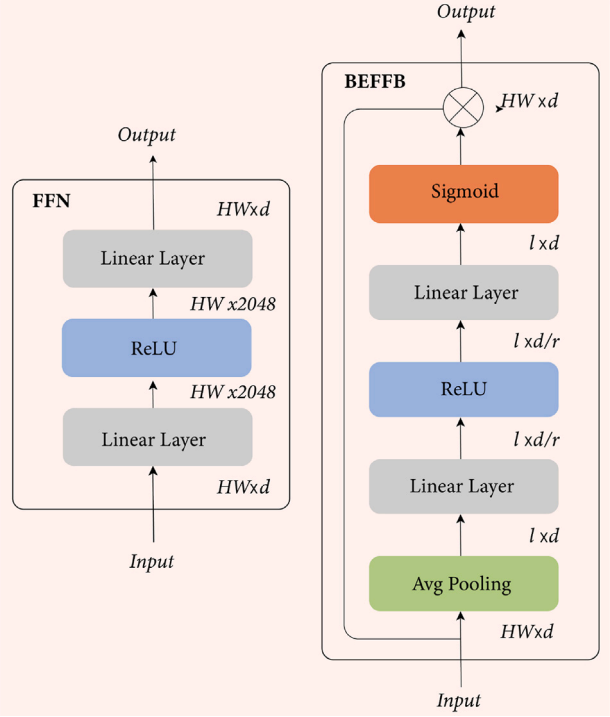


Fig. 3. The architecture of the vanilla FFN in the transformer [71] and our proposed BEFFB.

multi-head self-attention mechanisms, focus on unique semantic contexts within each modality and adaptively enhance the feature representations in their respective modal branches. To ensure efficient use of computational resources and model compactness, the SAMRs in each modality fusion layer of the fusion transformer share weights, significantly reducing the number of parameters involved in the process as used in [87]. The refined feature representations obtained from SAMR then pass through two distinct paths in the CAMIs as query, key, and value inputs. Unimodal features from the SAMRs interact and fuse in the CAMIs, which harness the power of multi-head cross-attention, enabling the interaction and fusion of information derived from the two modalities. As illustrated with a blue box in Fig. 2, the fusion transformer is constructed by stacking L cross-modality fusion layers, each containing two SAMRs and CAMIs. As a result, the model capitalizes on the inherent complementary relationships between RGB and depth data.

3.4. Bottleneck enhancement feed-forward block

Attention mechanisms play a pivotal role in capturing and enhancing salient features. However, these mechanisms inherently possess limited fitting capacity. In the vanilla transformer architecture [71], both the encoder and decoder units incorporate the FFN to amplify their representational abilities. While the FFN boosts the overall performance, it comes at the expense of a significant increase in the model's parameters, leading to potential issues related to computational resources and model overfitting. To balance the need for enhanced representational power and the practical constraints of maintaining a manageable number of model parameters, we propose the bottleneck enhancement feed-forward block (BEFFB). This proposal is inspired by two key architectural designs: (1) the bottleneck concept derived from ResNet [75], and (2) the squeeze-and-excitation block [88]. BEFFB is designed to augment the model's expressive capability while concurrently streamlining its parameter count. This is achieved by using a bottleneck structure that shrinks and then expands the dimensionality through ResNet and a channel-wise re-calibration step inspired by the SE block that focuses on enhancing the most salient features. Fig. 3 showcases the structural dissimilarities between the traditional FFN and the proposed BEFFB.

Given an input feature $X \in R^{b \times HW \times d}$ for BEFFB, where b represents the batch size, HW denotes the sequence length, and d is the embedding dimension, where embedding is a transformation of high-dimensional data into a lower-dimensional space. Initially, to capture the global context of the input features, we compress the sequence information utilizing global average pooling. This operation creates a global context embedding, z, which captures the average information across all spatial locations for each feature channel. This embedded representation encapsulates the global contextual information of the input feature sequence and serves as an essential element in the following feature enhancement process. The nth element of z is determined as:

$$z(n) = \frac{1}{HW} \sum_{i}^{HW} X(i, n) \tag{1}$$

Next, we exploit the global context embedding z, generated from the squeeze operation, to re-calibrate the input features across the embedding dimension d. This step scales and modifies the initial input features based on the global contextual information in BEFFB. The result is a more context-aware and refined feature representation, enhancing the model's ability to focus on relevant information. In particular, the re-calibration process is achieved by deploying a pair of dense layers with non-linearity, forming a bottleneck structure. This structure has a lower number of parameters as compared to FNN and helps to learn more complex and discriminative feature representations. Furthermore, the nonlinearity ensures the model captures non-linear dependencies across the features, making the transformation more robust. This procedure can be described as follows:

$$\sigma = sigmoid(W_2 ReLU(W_1 z)) \tag{2}$$

where $\sigma \in R^{N \times 1 \times d}$ the excitation value which indicates the significance of each feature channel, with $W_1 \in R^{\frac{d}{r} \times d}$ and $W_2 \in R^{d \times \frac{d}{r}}$. Rectified linear activation function (ReLU) and sigmoid (σ) are incorporated to inject nonlinearity into the transformations, thus enabling the model to capture more complex relationships within the data. The two dense layers, which include a reduction ratio of r control the complexity. The impact of r is further explored in Section 4.4.3. Finally, the block's output is generated by recalibrating the input feature using the bottleneck activation.

$$x = x_{HW} \otimes \sigma \tag{3}$$

where $\tilde{X} = [x_1, x_2, x_3, \ldots, x_{HW}]$ and $\otimes$ represent the Hadamard product. This operation multiplies each element of the input feature map X with the corresponding element in the excitation value σ, essentially scaling each feature vector with its corresponding importance value. This re-calibration operation is crucial in enabling the BEFFB to either emphasize or suppress distinct elements within each feature vector in the sequence embedding. When compared to the FFN, the BEFFB has a lower number of parameters and reduced computational costs.

3.5. Feature fusion

To fully exploit the strengths of both unimodal and multimodal features, our fusion transformer incorporates two SAMRs and two CAMIs into a single fusion layer. As depicted in Fig. 4, the basic fusion layer is designed with a dual purpose. First, it employs the SAMRs to amplify the unique contextual characteristics within each modality. This module operates independently for each modality, enhancing the distinguishing features and refining the representation. Second, the basic fusion layer utilizes the CAMIs to promote intercommunication between the different modalities. The CAMIs facilitate the interaction between the RGB and depth modalities, enabling the model to exploit

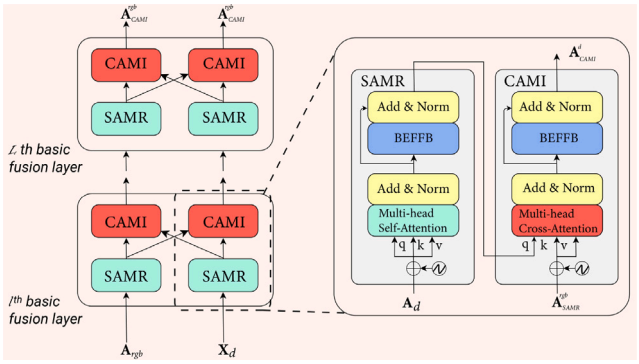


Fig. 4. The architecture of our fusion transformer where each cross-modality fusion layer consists of two SAMRs and two CAMIs.

the complementary and distinctive features inherent in each modality. The fusion transformer is comprised of L basic fusion layers stacked together, which can efficiently learn multimodal features by alternating between SAMR-based refinement and CAMI-based interaction. The SAMR and CAMI are described in more detail below.

3.5.1. Self-attentive modality refinement module

As illustrated in Fig. 4, SAMR encompasses two key components: multi-head self-attention, two normalization layers, and a BEFFB. The attention mechanism is a fundamental part of our CMOT. The rationale behind this lies in the inherent ability of the attention mechanism to effectively model dependencies in the input features, thereby leading to a more comprehensive and rich feature representation. The use of multiple attention heads [71], further enriches this attention by focusing on diverse aspects of the input. The multi-head attention mechanism, given queries Q, keys K, and values V, can be described as follows:

$$MultiHead(Q,K,V) = Concat(h_1,h_2,\dots,h_n)W \tag{4}$$

$$h_i = \text{Attn}(QW_i^Q, KW_i^K, VW_i^V) = \text{softmax}\left(\frac{QK}{\sqrt{d_k}}\right)V \tag{5}$$

where the key dimension is d_k, and the parameter matrices are $W_i^Q \in R^{d_n\times d_k}$, $W_i^K \in R^{d_n\times d_k}$, $W_i^V \in R^{d_n\times d_v}$, and $W_i \in R^{hd_n\times d_n}$. By default, we set h = 8, $d_m = 1024$, and $d_k = d_v = d_n/\text{h} = 64$. The SAMR module is specifically designed to enhance the semantic richness of the features by including multi-head self-attention and bottleneck enhancement. In our fusion transformer, the SAMRs in each layer share weights to reduce the parameter count. However, it is worth noting that the attention mechanism described in Eq. (5) lacks the ability to differentiate the positional information present in the input feature sequence. To mitigate this, we incorporate a spatial positional encoding leveraging a sine function on the input, as proposed in [88]. This creates an embedded attention feature that can establish interdependencies and facilitate their re-calibration through the ensuing BEFFB. Furthermore, to enhance optimization, we employ a residual connection [75] coupled with a layer normalization (LN) function [89] to alleviate the internal covariant shift and aid in preserving the network's stability. This pairing supports a smoother and more effective learning process, enhancing the robustness and stability of the model. The operational process of the SAMR can be detailed as follows:

$$A_{SAMR} = LN(\tilde{A}_{SAMR} + BEFFB(\tilde{A}_{SAMR})) \tag{6}$$

$$\tilde{A}_{SAMR} = LN(A + MultiHead(\tilde{A},\tilde{A},\tilde{A})) \tag{7}$$

$$\tilde{A} = A + P_A \tag{8}$$

where $A \in R^{N\times HW\times d}$ represents the input feature of a specific modality, $P_A \in R^{N\times HW\times d}$ denotes the positional encoding vector, and $A_{SAMR} \in R^{N\times HW\times d}$ corresponds to the SAMR output. Through this procedure, the SAMR module effectively refines the initial feature by leveraging spatial positional encoding and multi-head self-attention, thereby enhancing its semantically relevant elements and providing a more context-aware feature representation.

3.5.2. Cross-attentive modality interaction module

The CAMI structure, as illustrated in Fig. 4, comprises a multi-head cross-attention, two normalization layers, and a BEFFB. The CAMI accepts input features from both streams and concurrently processes their features for multi-modal fusion. The representation capability of these features, combined through multi-head cross-attention, is enhanced by their respective BEFFBs. Ultimately, the CAMI features of both branches are combined to produce the final output. This entire procedure can be described as:

$$A_{fusion} = A^r_{CAMI}gb + A^d_{CAMI} \tag{9}$$

$$A_{CAMI} = LN(\tilde{A}_{CAMI} + BEFFB(\tilde{A}_{CAMI})) \tag{10}$$

$$A_{CAMI} = LN(A_q + MultiHead(\tilde{A}_q,\tilde{A}_{kv},\tilde{A}_{kv})) \tag{11}$$

$$\tilde{A}_q = A_q + P_q \tag{12}$$

$$\tilde{A}_{kv} = A_{kv} + P_{kv} \tag{13}$$

In this case, $A_q \in R^{N\times HW\times d}$ represents the input of the branch where the CAMI is used, and $P_q \in R^{N\times HW\times d}$ corresponds to the spatial positional encoding associated with A_q. The input from another branch is denoted by $A_q.A_{kv} \in R^{N\times HW\times d}$, and its spatial encoding is represented by $P_{kv} \in R^{N\times HW\times d}$. The CAMI output is $A_{kv}.A_{CAMI} \in R^{N\times HW\times d}$, while the final $A_{fusion} \in R^{N\times HW\times d}$ is derived by combining the CAMI features from both the RGB and depth streams.

3.6. Modality-level triplet loss

For the online learning setting within the CMOT, we reconfigure the modality-level triplet loss to align with the principle of distance learning, which inherently suits the dynamic nature of online learning. The fundamental shift in this scenario arises from the fact that we are not dealing with a static set of classes. Instead, new classes, in the form of unseen individuals, continue to emerge during the learning process. Consequently, we need to transition from conventional pairwise comparisons, typical in triplet loss, to a more flexible mechanism that computes a distribution over all classes for each query sample. In order to make the model discriminative, we propose a modality-level triplet loss that focuses on increasing the distance between dissimilar instances while decreasing the distance between similar instances in the embedding space. This loss function operates on the extracted feature representations from the three main streams of our CMOT: $A_{rgb} \in R^{HW\times d}$ features from the dynamic RGB image, $A_d \in R^{HW\times d}$ dynamic depth image, and $A_{fusion} \in R^{HW\times d}$ from SAMR and CAMI. Consequently, we need to transition from conventional pairwise comparisons to a more flexible mechanism that computes a distribution over all classes for each query sample. The individual modality loss functions are defined as follows:

$$L_{rgb} = \frac{-log(exp(-D(A_{rgb\_q},A_{rgb\_k})))}{\sum_{\hat{k}}(exp(-D(A_{rgb\_q},A_{rgb\_\hat{k}})))} \tag{14}$$

$$L_d = \frac{-log(exp(-D(A_{d\_q},A_{d\_k})))}{\sum_{\hat{k}}(exp(-D(A_{d\_q},A_{d\_\hat{k}})))} \tag{15}$$

$$L_{fusion} = \frac{-log(exp(-D(A_{fusion\_q},A_{fusion\_k})))}{\sum_{\hat{k}}(exp(-D(A_{fusion\_q},A_{fusion\_\hat{k}})))} \tag{16}$$

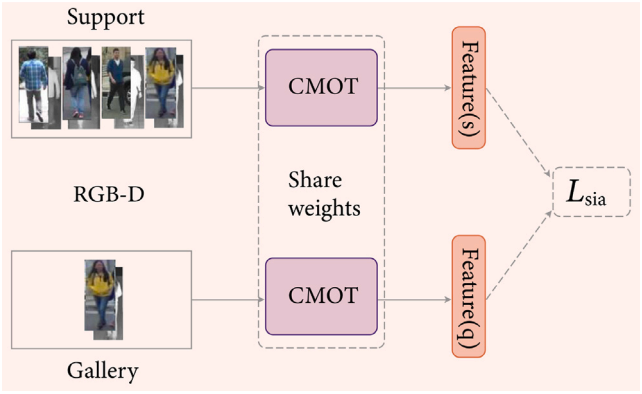


Fig. 5. The Siamese network takes a pair of samples as its input. The weights are shared across the two branches. The Siamese loss and identity loss are jointly implemented on the features learned from the support and gallery.

where A_{*_q} stands the respective modality query feature vector which represents the unique set of characteristics derived from an individual currently under consideration. A_{*_k} denotes the class descriptor, the central vector of the feature representations corresponding to a known class. In contrast, $A_{*\_\hat{k}}$ refers to the feature vectors of all other classes in the system. The distance function D(.) measures the cosine similarity, between these feature vectors. The lower the distance, the higher the similarity. The term -log(exp(-D(A_{*_q}, A_{*_k}))) computes the negative log-likelihood of the correct class, given the query sample. The denominator, $\sum_{\hat{k}}(exp(-D(A_{*\_q}, A_{*\_\hat{k}})))$, normalizes this likelihood overall classes, ensuring that the probabilities add up to 1. This computation essentially represents a softmax function over the distances from the query sample to all class descriptors.

Next, we incorporate a balancing parameter $\lambda \in [0, 1]$ to control the contributions of unimodal feature extraction losses (L_{rgb} and L_d) and the multi-modal fusion loss L_{fusion}. The unimodal losses emphasize the individual aspects of the RGB and Depth modalities, whereas the fusion loss brings them together in a unified representation. When λ is close to 1, the model puts more emphasis on unimodal features, and when λ is close to 0, it emphasizes the fusion features more. The final modality-level triplet loss $L_{triplet}$ is calculated as:

$$L_{triplet} = \lambda(L_{rgb} + L_d) + (1 - \lambda)L_{fusion} \tag{17}$$

3.7. Online learning

We have approached online learning as a K-way C-shot few-shot learning task, where our computational model aims to quickly adapt and classify new individuals by leveraging a limited number of labelled examples (C samples) for each identity that is not part of the initial training data. Our dataset is denoted as $G = \{(x_1, y_1), \ldots, (x_N, y_N)\}$, where N represents the total number of person's IDs and x is the training sample with its associated label ID y. Given a specific K-way C-shot meta-task T, the identity labels are selected at random from the set G. These labels are symbolized by $V = \{y_1 | i = 1, \ldots, K\}$. Consequent to this selection, training examples from these selected identities are subsequently harnessed to create two distinct sets: a support set (S) and a query set (Q).

1. The support set for task T, denoted as S, comprises $C \times K$ samples (K-way C-shot).
2. The query set, symbolized as Q, includes n samples chosen for meta-testing.

We introduce softmax regression into the final layers of three streams for predicting person identities and incorporate a dual-branch weight-shared Siamese network that is designed to pull identical instance features closer while pushing different person features apart, as shown in Fig. 5. It is important to underscore that the Siamese network is tailored to learn similarity at the instance level, making it distinct from the proposed loss at the region level. We have approached online learning as a K-way C-shot few-shot learning task, where our computational model aims to quickly adapt and classify new individuals by leveraging a limited number of labelled examples (C samples) for each identity that is not part of the initial training data. Given an RGB-D pair of a person, s, and q, their final full-stream features are denoted as A(s) and A(q), respectively. Following this, the corresponding loss for the Siamese network can be defined accordingly.

$$L_{siamese} = \begin{bmatrix} \|A(s) - A(q)\|_2^2 & s = q \\ max\{(m - \|A(s) - A(q)\|_2^2), 0\} & s \neq q \end{bmatrix} \tag{18}$$

where m is a margin factor that is empirically set to 8 in our experiments. Following this [41], we jointly train the CMOT under triplet loss and siamese loss to boost person reID performance further. Taking into account the modality-level loss in CMOT, the cumulative loss for a pair of samples, specifically (s, q), can be denoted in the following manner:

$$Loss_{cumulative} = L_{triplet} + Loss_{siamese} \tag{19}$$

In reID, the primary objective is to match and identify individuals across different scenarios or camera views, while in K-way C-shot few-shot learning, the primary objective is to classify new instances into one of the K predefined categories, given only C examples per category. We have formulated pedestrian reID as a K-way C-shot few-shot learning task to address the challenges of adapting to new individuals quickly. This formulation allows us to leverage a limited number of labelled examples for each identity not present in the initial training data. However, it is important to emphasize that our task is distinct from traditional few-shot learning tasks because we are working with person reID, where the goal is to match individuals based on their RGB-D features.

4. Experiment and results

In order to evaluate the effectiveness of our CMOT, we conducted experiments on three prominent RGB-D person reID datasets: BIWI RGBD-ID [46], RobotPKU RGBD-ID [47] and TVPR2 [3] dataset, their summary is given in Table 1. These datasets have been selected for our experimental evaluation due to their relatively large sizes, varying sensor viewpoints, and complex environmental conditions, allowing for a more robust assessment of our CMOT performance. To analyse the contribution of each component in our model, we performed comprehensive ablation studies. These studies involved the evaluation of different components, such as fusion techniques, BRFFB, and different backbone networks, of our proposed method to assess their individual impacts on the overall performance. Furthermore, in order to provide a fair comparison with existing state-of-the-art models, we adopted identical evaluation protocols and utilized optimal components for our proposed method.

4.1. Datasets

4.1.1. BIWI RGBD-I1

The BIWI RGBD-ID [46] dataset is a specialized collection of data intended for the long-term reID of individuals through the utilization of RGB-D cameras. Comprising 50 distinct subjects, the dataset is divided into 50 training and 56 testing sequences. Each component of the dataset, captured at the highest attainable resolution of 1280 × 960 pixels with a Microsoft Kinect for Windows, includes synchronized RGB images, depth images, segmentation maps of persons, skeletal data, and ground plane coordinates. The videos were acquired at an approximate

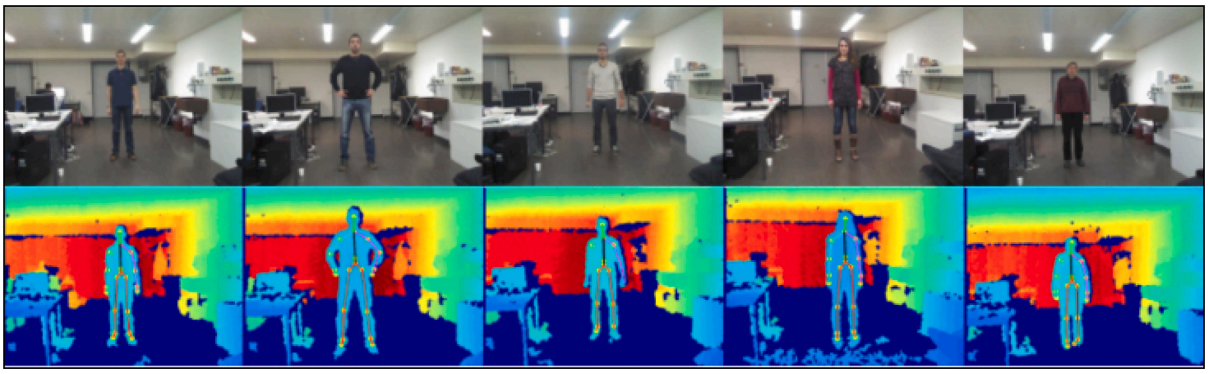


Fig. 6. BRGB and depth samples of the BIWI RGBD-ID dataset.

rate of 10 frames per second (fps). Within the training component, subjects are recorded performing specific routines of motion before the camera, including rotations around the vertical axis, various head movements, and two frontal walks. Of the 50 subjects in the training set, 28 were also recorded in two testing videos each. These testing sequences were collected on different days and in different locations compared to the training dataset, resulting in variations in attire among the subjects. Each individual in the testing set was recorded in two distinct sequences: a "still" sequence and a "walking" sequence. In the former, subjects remain largely motionless or exhibit slight movement in place, whereas the latter features each subject performing two frontal and two diagonal walks relative to the Kinect.

The BIWI RGBD-ID dataset was prepared with the specific objective of furnishing a valuable resource for reID predicated solely on body cues. In the "still" sequences, subjects are positioned at a distance of approximately 2–3 metres from the camera, as evidenced in Fig. 6, and in the "walking" sequences, although moving closer, they maintain a distance greater than approximately 1.5 m, often providing side views of the face. Within this context, meaningful reconstructions of facial features are unattainable. This particular characteristic has generated two significant implications for the experimental evaluation. Firstly, the BIWI dataset was employed to assess the efficacy of the skeletal component alone within our reID model, compared with contemporary state-of-the-art reID models. Second, it is crucial to mention that the current literature lacks an RGB-D reID dataset that facilitates the simultaneous analysis of both body and facial features.

4.1.2. RobotPKURGBD-ID

RobotPKU RGBD-ID [47] is captured by employing Kinect sensors in conjunction with the Microsoft Kinect SDK. It comprises a total of 180 video sequences captured from 90 individuals. Each individual's video sequences encompass both static (still) and dynamic (walking) sequences, which were recorded in two distinct indoor locations. However, it is important to note that certain video sequences within the dataset may exhibit inherent limitations, such as noisy depth frames or incomplete body parts in the captured images. Such limitations can arise when depth sensor-based cameras capture images outside their optimal range or encounter challenges in accurately capturing all body parts. To mitigate the impact of these limitations, preprocessing techniques introduced in a prior study [26] were employed. In line with these pre-processing techniques, any frames deemed improper or unreliable were discarded from the dataset. This selection process ensured that only RGB frames accompanied by corresponding accurate depth images, providing a comprehensive representation of an individual, are considered for the experimental analysis.

4.1.3. TVPR2

The TVPR2 [3] is specifically designed for top-view reID research. It consists of 200 videos captured in a hallway, featuring a total of 1000 individuals, with each individual appearing twice. This dataset is divided into a train set and a test set, where individuals walking from left to right are included in the training set, and individuals walking from right to left form the test set. For our evaluation, we have focused on a subset of 94 individuals due to missing frames in one of the recorded videos. TVPR2 captures depth frames in a top-view configuration, with a resolution of 320 × 240 pixels. The depth frames were acquired using an Asus Xtion Pro Live camera, which was ceiling-mounted to provide an unobstructed view of individuals passing beneath it. This setup was chosen to minimize occlusions and prioritize privacy, as facial images are not recorded. The recordings took place over the course of eight days in an indoor environment, capturing variations in lighting conditions that are influenced by natural light throughout the day. The participants in the dataset range in age from 19 to 36 and represent diverse ethnic backgrounds. They were dressed in everyday clothing such as t-shirts, sweatshirts, shirts, loose pants, coats, scarves, and hats, reflecting real-world scenarios. Notably, the dataset includes three pairs of twins, further diversifying the dataset's composition. All videos in the TVPR2 dataset have fixed dimensions of 320 × 240 pixels and were recorded at a frame rate of approximately 30 frames per second. Manual annotations are provided, indicating the recording time, the number of individuals in each session, and the order in which individuals pass the camera, facilitating accurate analysis and evaluation. The TVPR2 dataset offers a realistic and challenging environment for top-view person re-identification studies, encompassing variations in lighting, diverse clothing appearances, and occlusion-free depth frames.

Table 1
Summary of all the used datasets.

Dataset	BIWI RGBD-ID	RobotPKU RGBD-ID	TVPR
Sensor	Kinect V1	Kinect	Asus Xtion Pro Live
Dataset Size	106 sequence	180 video	235 video
Total Individuals	50	90	1027
Data Modality	RGB-D	RGB-D	RGB-D
SDK Used	Microsoft SDK	Microsoft SDK	PrimeSense SDK

4.2. Implementation setting

4.2.1. Network inputs

Hypothesis generation module, which uses an object detector, Yolov5 [37] and StrogSORT tracker [83], creates the sequence of frames that contain precisely cropped region-of-interest of the person both from RGB and depth modality. Different from the conventional two-stream framework [49,60], which primarily focuses on the fusion of multi-modality through the concatenation of extracted feature vectors, our proposed approach incorporates a dual-stream feature extraction CNN to process the hypothesis that jointly learns multi-modal features using pairs of dynamic RGB and Depth images, denoted as $\langle VDS, DDS \rangle$. For training purposes, the input RGB-D sequences are

resized to 256 × 256 pixels. To facilitate further analysis, the depth images are converted into greyscale images in the [0, 255] range. Image augmentation techniques are employed to enhance dataset diversity and network performance for both the RGB and RGB-D models. Augmentation techniques [60], including random cropping and flipping are applied during the training process. Furthermore, all images are resized to a fixed size of 256 × 256 pixels to ensure consistency and compatibility across the models. By incorporating dual-stream feature extraction CNNs and employing data augmentation techniques, our approach aims to capture rich multi-modal features while accounting for variations in the input RGB-D sequences.

4.2.2. Training setting

In our CMOT, we initialize the feature extraction CNN using pre-trained weights from the MS-COCO [90] dataset. This pre-training serves as a weight initialization step for the feature extraction process. Our proposed method is implemented using PyTorch, with training and validation executed on a machine with a 16-core CPU @ 3.60 GHz, supported by an Nvidia GeForce RTX 4090 GPU with 128 GB RAM. To optimize the model parameters, we employ the Adam optimizer. The training process continues for 150 epochs with a batch size of 32. A momentum factor of 0.8 is applied to enhance the optimization process's stability and convergence. The initial learning rate for the CNN is set to 0.02. We also incorporate a learning rate decay strategy for stabilizing the training process, reducing the learning rate by a factor of 0.1 every 25 epochs. Similarly, the transformer component within our framework has an initial learning rate of 0.01. The transformer's learning rate follows a decay schedule, decreasing by a factor of 0.1 at the 30th and 50th epochs. Based on our observations during experimentation, we find that early stopping improves the results on the BIWI RGBD-ID dataset. Consequently, we utilize the model's performance at the 115th epoch for comparative evaluation and analysis.

4.2.3. Evaluation protocol

Our evaluation uses standard person reID evaluation measures, specifically the Cumulative Matching Characteristic (CMC) Rank and Mean Average Precision (mAP) [9,49,60]. The dataset is evenly divided into training and testing subsets to guarantee impartial evaluation. In the test phase, each RGB-D query image is contrasted against RGB-D gallery images. The trained network's feature embeddings are utilized to measure the similarity between each query image and the entire gallery set. This dissimilarity is measured using the cosine distance, resulting in a vector of dissimilarity scores. To combine the scores from both RGB and RGB-D modalities, we merge them in the dissimilarity space. Our reID CMOT then generates a ranked list of the top n images with the lowest dissimilarity to the query image from the gallery set. The performance is evaluated based on the rank at which the query image's true match appears in the list. Specifically, if the true match is found at the kth position, the query is considered to have achieved rank k. To ensure robustness, we repeat the experiments 10 times and report the average accuracies for ranks 1, 5, and 10, along with the mAP. It is important to emphasize that all results presented in this paper are obtained under the single query setting, where each query image is evaluated independently. This setting ensures fairness and consistency in the comparison of different approaches and allows for a comprehensive evaluation of the proposed method's performance. By employing the CMC-rank and mAP, we aim to provide a thorough and reliable assessment of the effectiveness and accuracy of our approach for person re-identification tasks. These evaluation metrics offer insights into both the ranking performance and the overall quality of the retrieved results, enabling a comprehensive analysis of the proposed method's capabilities.

Table 2
Comparison of our CMOT on RGB, Depth, and RGB-D on all three datasets.

Dataset	Modality	Rank-1	Rank-5	Rank-10	Rank-20	mAP
BIWI RGBD-ID	RGB	73.51	84.42	89.71	92.71	73.52
	Depth	64.80	74.48	80.7	86.56	67.71
	RGB-Depth	77.84	91.25	97.32	98.77	77.61
RobotPKU RGBD-ID	RGB	90.80	92.09	94.64	95.51	86.94
	Depth	91.70	92.70	94.82	96.81	87.64
	RGB-Depth	92.10	94.51	95.52	98.52	90.24
TVPR2	RGB	95.21	97.64	98.10	98.48	81.44
	Depth	94.50	95.17	97.81	98.08	80.11
	RGB-Depth	99.1	99.13	99.14	99.99	84.21

4.3. CMOT evaluation

We present the results of our experiments conducted on three datasets to demonstrate the effectiveness of our proposed similarity-based self and cross-attention cross-modality fusion model on each of these datasets. We first provide a comprehensive analysis and performance comparison of our CMOT. We evaluate the accuracy and robustness of our model by measuring various metrics, such as rank-1, rank-5, rank-10, and mAP. Moreover, we also compare our CMOT against the proposed cross-modality reID techniques that have been previously applied to the corresponding datasets.

4.3.1. Comparison on datasets

This experiment compares the performance of our CMOT for RGB, depth, and RGD-D modalities on multiple datasets with varying levels of complexity arising from factors such as scene setting, lighting conditions, occlusion, and visual perception. The validation set of BIWI RGBD-ID, RobotPKU RGBD-ID, and TVPR2 datasets are utilized for this purpose, and the results are presented in Fig. 7 in the form of CMC curves. The CMC curves provide insights into the rank-based performance of our fusion model across different datasets. It is observed that the proposed CMOT model achieves the highest rank-1 on the TVPR2 dataset, indicating excellent performance in correctly identifying the top match. Conversely, the BIWI RGBD-ID dataset exhibits the lowest rank-1 accuracy, highlighting its inherent complexity. As shown in Table 2, in the BIWI RGBD-ID dataset, the CMOT model displays substantial improvements when shifting from RGB and Depth modalities to RGB-Depth. The rank-1 percentage improved from 73.51% and 64.8% to 77.84% for RGB, depth, and RGB-depth modalities, respectively, showing a noticeable increase of 5.87% and 20.01% from the RGB and Depth modalities. Similarly, the mAP values also increased from 73.52% and 67.71% to 77.61%, recording an improvement of 5.56% and 14.63% respectively. For the RobotPKU RGBD-ID dataset, the model improvements are not as drastic, but still apparent. The rank-1 values increase from 90.8% and 91.7% to 92.1% for RGB, Depth, and RGB-Depth modalities, marking a 1.43% and 0.44% rise. The mAP values show a similar trend, increasing from 86.94% and 87.64% to 90.24%, reflecting a 3.79% and 2.97% improvement, respectively. Finally, in the TVPR2 dataset, the CMOT model's improvements from RGB and Depth to RGB-D are substantial. The rank-1 percentage improved from 95.21% and 94.50% to 99.1%, showing a stark increase of 4.09% and 4.86% from the RGB and Depth modalities. Meanwhile, the mAP values also increased from 81.44% and 80.11% to 84.21%, recording an improvement of 3.4% and 5.12% respectively. Overall, the findings demonstrate the effectiveness and adaptability of the CMOT model in handling RGB-D reID tasks across diverse datasets, and it is evident that the combined RGB-Depth modality considerably improves the CMOT model's performance across all datasets.

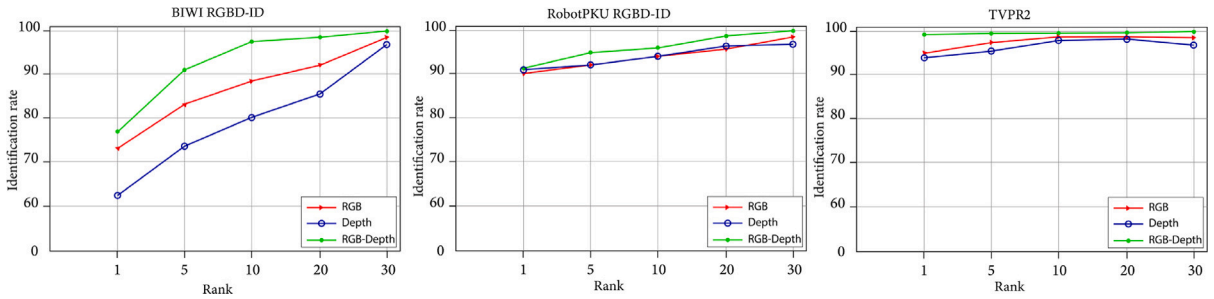


Fig. 7. BIWI RGBD-ID, RobotPKU REGD-ID, and TVPR2 dataset respectively for RGB, depth, and RGB-Depth modality.

4.3.2. Comparison with state-of-the-art reID methods

BIWI RGBD-ID dataset. Table 3 shows the superior overall performance of our proposed approach on the BIWI RGBD-ID dataset compared to previous methods, highlighting the effectiveness of aligning dual-granularity features. Among the methods, CMOT emerges as a notable proposition, achieving an impressive 77.8% in rank-1 accuracy. The closest competitor in this regard is HRN [14], which attains 47.1% accuracy, thereby revealing a substantial increase of approximately 30.7% when compared with HRN. Other methods, such as SM-SGE [91] (34.8%), Distillation [92] (40.4%), SimMC [33] (41.7%), and Hi-MPC [93] (47.5%), fall behind CMOT in terms of rank-1 accuracy, with decrements ranging from 30.3% to 43.0%. In rank-5 accuracy, CMOT continues to demonstrate superiority, achieving 91.2%, while the nearest competitor, HRN, achieves 73.5%, showing an improvement of about 17.7% relative to HRN. CMOT also outperforms SM-SGE (60.6%), Distillation (77.1%), SimMC (66.6%), and Hi-MPC (70.3%) by substantial margins, with increments ranging from 30.6% to 20.9%. With regard to rank-10 accuracy, CMOT showcases exceptional performance at 97.3%, illustrating its ability to identify subjects correctly even within the top 10 matches. In contrast, HRN achieves 78.1% in rank-10 accuracy, indicating a significant increment of approximately 19.2% relative to HRN. Other methods, including SM-SGE (71.5%), Distillation (91.0%), SimMC (76.8%), and Hi-MPC (78.6%), are outperformed by CMOT, with increments ranging from 25.8% to 18.7%. In the context of the mAP metric, which evaluates the overall quality of ranking, CMOT further accentuates its effectiveness, reaching an mAP of 77.6%. The next closest method is HRN, with an mAP of 44.6%, representing an improvement of around 33% relative to HRN. Additional methods, such as SM-SGE (12.8%), Distillation (41.3%), SimMC (12.3%), and Hi-MPC (17.4%), manifest significantly lower mAP values, with increments ranging from 64.8% to 60.2%. In comparison to Distillation [92], which leverages dual-granularity features and similarity inference mechanisms to extract contextual relationships, our CMOT model demonstrates a more effective learning of modality-sharable features. By employing a cross-attention mechanism, our CMOT model can identify nuanced information across different levels of granularity, from coarse to fine. These results highlight the robustness of our method across different evaluation modes and further emphasize the effectiveness of our approach in learning and leveraging modality-sharable features.

Our CMOT demonstrates superiority over other state-of-the-art reID models attributed to several aspects of its design and components. The model's advanced feature extraction and effective integration of RGB and depth modalities enable the comprehensive feature representations, crucial for reID tasks. CMOT's fusion transformer plays a pivotal role in refining and optimizing the fusion of modalities, allowing the model to exploit distinctive and complementary features effectively, yielding highly discriminative and robust representations. Other reID depends on early fusion or late feature fusion techniques, however, these techniques do not leverage the cross-modality discriminative attributes in RGB and depth modality. The model's capability to adapt in an online learning setting allows continual refinement and adaptation to new data points, which is essential in dynamic environments characterized by the influx of unseen individuals. The implementation of modality-level triplet loss in CMOT enhances its discriminative power which combines the RGB, depth, and RGB-depth modality losses, enhancing accuracy in reID tasks.

Table 3
Results of our CMOT model and SOTA reID approaches one BIWI RGBD-ID dataset.

Method	Rank-1	Rank-5	Rank-10	mAP	Online learning
SM-SGE [91]	34.8	60.6	71.5	12.8	No
HRN [14]	47.1	73.5	78.1	44.6	No
Distillation [92]	40.4	77.1	91.0	41.3	Yes
SimMC [33]	41.7	66.6	76.83	12.3	No
Hi-MPC [93]	47.5	70.3	78.6	17.4	No
CMOT (ours)	77.8	91.2	97.3	77.6	Yes

BRobotPKU dataset. Table 4 indicates that the performance on the RobotPKU dataset where our CMOT achieves a rank-1 of 92.1%, which is significantly higher than that of FFM [11], by 14.2%. However, it is marginally lower than depth-guided attention [94] by 0.1% and dissimilarity fusion [60] by 1.2%. In terms of rank-10, the CMOT model registers a score of 95.52%, which is slightly lower than dissimilarity fusion by 0.48%. The CMOT model excels in rank-20 accuracy, surpassing all other methods with a score of 98.52%. This performance constitutes a substantial improvement of 1.92% compared to dissimilarity fusion [60], which has a rank-20 accuracy of 96.6%. Boasting the highest mAP score of 90.24%, the CMOT model exhibits a 0.84% improvement over dissimilarity fusion, which has an mAP score of 89.4%. The table does not provide mAP scores for FFM and depth-guided attention. While the CMOT model outperforms some methods in certain performance metrics, it shows slightly lower performance in others. Overall, the CMOT model's effectiveness and robustness in handling RGB-D reID on the RobotPKU dataset are evident.

TVPR2 dataset. Table 5 showcases a comparison between CMOT and various SOTA methods on the TVPR2 dataset. CMOT achieves a rank-1 of 99.1%, which is higher than all other methods except for SeSAME [25] at 98.8%, indicating a 0.3% improvement in rank-1. For rank-5, the CMOT model attains a score of 99.14%, which is lower than SeSAME at 99.4% by 0.26%. The CMOT model achieves the highest rank-10 accuracy of 99.99%, outperforming all other methods, including SeSAME with a rank-10 accuracy of 99.7%. In terms of mAP scores, the CMOT model demonstrates the best performance with a score of 84.21%, surpassing SeSAME by 2.61%, which has an mAP score of 81.6%. In conclusion, the CMOT model exhibits competitive performance on the TVPR2 dataset compared to other SOTA methods, with improvements in rank-1, rank-10, and mAP scores and a slight decrease in rank-5 accuracy compared to SeSAME.

The result comparison of all these datasets shows different trends. The comparison on the BIWI RGBD-ID dataset illustrates the great performance improvement when compared to other models that showed low rank-1 and mAP as compared to previous models that showed performance on the other two datasets, RobotPKU and TVPR2. This shows that the BIWI RGBD-ID dataset is inherently complex dataset due to the involvement of indoor and outdoor environments, which

Table 4
Results for our CMOT and SOTA reID approaches on the RobotPKU dataset.

Method	Rank-1	Rank-10	Rank-20	mAP	Online learning
FFM [11]	77.9				Yes
Depth guided attention [94]	92.0				No
Dissimilarity fusion [60]	93.3	96.0	96.6	89.4	Yes
CMOT (ours)	92.1	95.52	98.52	90.24	Yes

Table 5
The experimental results for our CMOT and SOTA reID approaches on the TVPR2 dataset.

Method	Rank-1	Rank-5	Rank-10	mAP	Online learning
VRAI-Net3 [9]	74.48	–	–	7794	Yes
RGB-D-CNN [11]	81.4	85.2	91.1	70.7	No
MAT [94]	91.1	93.7	94.1	78.8	No
SLATT [95]	90.6	92.1	94.3	75.1	No
SeSAME [25]	98.8	99.4	99.7	81.6	Yes
CMOT (ours)	99.1	99.14	99.99	84.21	Yes

Table 6
Comparison of feature fusion techniques on the cross-modality.

Fusion technique	SAMR	CAMI	Rank-1	mAP
CM-Avg	–	–	69.81	67.32
CM-Mul	–	–	71.21	70.56
CM-Concat	–	–	72.08	71.81
$CMOT_{SAMR}$	√	–	75.21	75.50
$CMOT_{CAMI}$	–	√	76.40	77.02
CMOT	√	√	77.84	77.61

exhibit more variations in lighting, weather conditions, and camera angles compared to the other two datasets. Moreover, our CMOT has shown slightly inferior performance as compared to other models. This can be attributed to the data-capturing process which used robot-mounted depth sensors to collect the data. This setup can lead to specific challenges related to camera placement and mobility which are problematic for our CMOT model.

4.4. Ablation study

We have conducted ablation studies on the BIWI RGBD-ID dataset, known for its higher complexity, to evaluate different RGB-D feature fusion techniques including our fusion transformer and improved form of the FFN, BEFFB, Our experiments are conducted using a ResNet-50 [75] backbone network, where the final classification layer is removed and the network is initialized with pre-trained parameters from MS-COCO. These ablation studies allow us to analyse the contribution and impact of our proposed components on the overall performance of the CMOT [47].

4.4.1. Fusion technique

The effectiveness of the cross-modal fusion transformer is investigated by comparing its different structures with other popular fusion methods within a dual-stream framework. The following networks are examined and compared:

- CM-WA represents the fusion method based on the weighted average score which receives the feature vectors of both modalities extracted from dual-stream CNN.
- CM-M refers to a fusion technique where output score vectors of RGB and depth modalities are extracted from dual-stream CNNs fused through elementwise multiplication.
- CM-Concat fuses the features of RGB and depth modalities obtained from the dense layer which receives the concatenated feature maps from the dual stream as input.
- $CMOT_{SAMR}$ only utilizes the self-attention-driven SAMR in the cross-modality fusion transformer.
- $CMOT_{CAMI}$ retains cross-attention-driven CAMI while removing SAMR from the cross-modal fusion transformer.
- CMOT represents the SAMR and CAMI-based proposed reID framework as illustrated in Fig. 2

Table 6 presents the results comparing CMOT with other fusion methods, highlighting the effectiveness of the proposed fusion structure. The inclusion of both self-attention-driven SAMR and cross-attention CAMI within CMOT is found to be crucial for achieving successful multi-modal fusion. The comparison reveals that CMOT outperforms $CMOT_{SAMR}$ and $CMOT_{CAMI}$, indicating the effectiveness of extracting complementary information between uni-modal features by employing multiple basic fusion layers that alternate between SAMR and CAMI deployment. Notably, the basic fusion techniques, such as CM-WA, CM-M, and CM-Concat, exhibit lower performance compared to the advanced techniques utilizing SAMR and CAMI. Specifically, CM-Concat achieves a rank-1 of 72.08% and a mAP score of 71.81%, which, while higher than CM-WA and CM-M, are lower than the CMOT variants incorporating SAMR and CAMI. When SAMR is integrated into the CMOT model ($CMOT_{SAMR}$), a significant improvement in performance is observed, with a rank-1 of 75.21% and a mAP score of 75.50%. Further enhancements are achieved by adding CAMI to the CMOT model ($CMOT_{CAMI}$), resulting in a rank-1 of 76.40% and a mAP score of 77.02%. The highest performance is attained when both SAMR and CAMI are incorporated into the CMOT, yielding a rank-1 of 77.84% and a mAP score of 77.61%. These results demonstrate the superiority of advanced fusion techniques, such as SAMR and CAMI, in improving the performance of the CMOT model for the RGB-D reID task.

4.4.2. Effect of number of transformer layers

Following the vanilla vision-transformer [29] and evaluating the performance in-depth of our CMOT. We experiment with different numbers of cross-modality basic fusion layers ranging from 1 to 6 layers. The experiment outcome given in Table 7 reveals that deeper models tend to achieve better performance but at the cost of increased computational demands. As we increment the layer count denoted as "L", the performance of the CMOT model exhibits consistent enhancement and eventually reaches a plateau at L = 4. This plateau effect can be rationalized by the observation that an increase in the number of layers introduces instability in training gradients, making the optimization process more challenging. Such instability during training has been previously reported in [86,96]. Furthermore, it is noteworthy that the input to the CMOT consists of high-level features that have been extracted by CNNs. Given the richness of these high-level features, satisfactory results can be achieved without necessitating an excessively deep fusion network. Therefore, considering the interplay between reID performance and computational overhead, we opted to set L = 4 for our experimental configuration. This choice strikes a balance between achieving the reID performance and managing the computational resources efficiently.

4.4.3. Superiority of bottleneck excitation feed-forward

The integration of FFNs into the encoder and decoder components of a conventional transformer [72] has been shown to enhance the model's ability to represent complex information. In this study, we aim to investigate the impact of the hyperparameter r on the capacity and computational overhead of the BEFFB. The BEFFB is a module designed

Table 7
Comparison of the number of transformer layers.

Layers	Params (M)	FLOPs (M)	Rank-1	mAP
1	2.14	4.08	69.61	69.17
2	2.57	6.96	72.08	71.93
3	2.99	9.85	75.16	75.32
4	3.41	12.72	77.84	77.61
5	3.83	15.60	77.93	77.89
6	4.25	18.48	78.13	78.58

Table 8
Comparison of BEFFB with FFN and different reduction ratios.

Technique	Params (M)	FLOPs	Rank-1	mAP
CMOT-Non	–	–	73.32	67.32
CMOT-FFNA	19.84	0.68 G	75.61	70.56
CMOT-FFNB	19.84	39.41 M	77.99	77.04
$CMOT-BEFFB_1$	5.41	30.95 M	76.21	75.50
$CMOT-BEFFB_2$	4.32	24.51 M	76.40	75.02
$CMOT-BEFFB_4$	3.41	12.72 M	77.84	77.61
$CMOT-BEFFB_8$	1.52	4.98 M	75.99	74.91
$CMOT-BEFFB_{16}$	0.86	3.41 M	75.44	74.80

to reduce the parameters in the FFN while maintaining or even improving its performance. By introducing a bottleneck structure and an enhancement operation within the BEFFB, we aim to find a trade-off between model capacity and computational efficiency. The bottleneck architecture serves to compress the input features and capture the most salient information, while the enhancement operation further refines the representations by incorporating additional context or attention mechanisms. To evaluate the effectiveness of the BEFFB, we experimented with different configurations of the FNN and BEFFB. BEFFB configurations involve different values of the hyperparameter r which controls the reduction ratio of the bottleneck structure. By examining the performance of the model across these different configurations, we gain insights into the trade-off between capacity and computational requirements. The configurations of different feed-forward blocks are as follows:

- CMOT-Non: Our CMOT model without BEFFBs, serves as the baseline for comparison.
- CMOT-FFNA: Our CMOT model with vanilla FFNs replacing the BEFFBs.
- CMOT-FFNB: The CMOT model with FFNs, as well as global average pooling and excitation operations, is integrated into the CMOT-FFNA configuration.
- $CMOT-BEFFB_r$: The CMOT model with BEFFBs features a reduction ratio of r, where r is a hyperparameter determining the capacity and computational overhead of the BEFFBs.

Table 8 shows that performance does not consistently improve as the parameter r increases. CMOT-Non achieves a rank-1 of 73.32% and mAP of 67.32%. CMOT-FFNB performs the best with a rank-1 of 77.99% and mAP of 77.04%, exhibiting an improvement of 3.13% in rank-1 and 4.79% in mAP, compared to the baseline. This enhanced performance, however, comes with a considerable increase in parameters and FLOPs. CMOT-FFNB, the second variant of CMOT, further improves upon its performance. It achieves a rank-1 of 77.99% and a mAP of 77.04%, which translates to a 6.35% increase in rank-1 and a substantial 14.45% improvement in mAP when compared to the baseline. As the reduction ratio increases in the $CMOT-BEFFB_r$ configurations (from $CMOT-BEFFB_1$ to $CMOT-BEFFB_{16}$), there is a general trend of decreasing parameters and FLOPs, which indicates a reduction in computational complexity. For instance, $CMOT-BEFFB_1$ improves the rank-1 and mAP scores by 3.95% and 12.17% respectively, compared to the baseline. Moreover, it substantially reduces the parameters and FLOPs by 72.8% and 54.4% respectively when compared to CMOT-FFNB. Similarly, $CMOT-BEFFB_4$ also demonstrates impressive performance, achieving a rank-1 of 77.84% and an mAP of 77.61%. These scores signify an improvement of 6.16% in rank-1 and 15.29% in mAP over the baseline. Moreover, this model is quite resource-efficient, reducing the parameters by 82.9% and the FLOPs by 67.7% compared to CMOT-FFNB. However, as the reduction ratio increases (as in $CMOT-BEFFB_8$ to $CMOT-BEFFB_{16}$), there is a slight decline in performance compared to the models with lower reduction ratios. Despite this, these higher reduction ratio models still outperform the baseline. For example, $CMOT-BEFFB_{16}$ improves the rank-1 and mAP scores by 2.89% and 11.08% respectively, compared to the baseline. Simultaneously, it significantly reduces the parameters by 95.67% and the FLOPs by 91.35% when compared to CMOT-FFNB. A comparison of various BEFFB block configurations in the CMOT model reveals that selecting an appropriate reduction ratio is crucial for optimizing performance. While reducing the computational complexity is beneficial, it is essential to strike a balance between complexity and performance to achieve the best results in handling RGB-D reID. Consequently, we chose r = 4 for this study.

Table 9
Performance comparison on different modality losses.

Layers	Rank-1	Rank-5 (M)	Rank-10	mAP
L_{rgb}	62.73	88.27	92.59	58.93
L_d	57.48	87.64	91.77	51.60
L_{fusion}	74.31	90.14	94.82	73.55
$L_{triplet}$	77.84	91.25	97.32	77.61

4.4.4. Effectiveness of the triplet-modality loss

We assess the effectiveness of our $L_{triplet}$ loss with L_{rgb}, L_d and L_{fusion} when our CMOT is trained with each loss function and performance comparison is presented in the Table 9. First when CMOT is trained solely with the L_{rgb} loss function, it achieves a rank-1 of 62.73% and mAP of 58.93%. In contrast, training the CMOT exclusively with the L_d, which is based on depth modality, results in slightly lower performance across all ranks compared to L_{rgb} where 5.25% and 7.33% dip in rank-1 and mAP, respectively. This suggests that depth information alone is not as effective as RGB for person reID. The fusion of RGB and depth modalities using the L_{fusion} improves rank-1 to 74.31% and mAP to 73.55%. These results underscore the advantages of combining information from both modalities, demonstrating that the fusion approach is beneficial for person reID. Furthermore, the $L_{triplet}$ loss function, which integrates the L_{rgb}, L_d and L_{fusion} losses, emerges as the most effective loss function. The model achieves its highest performance with a rank-1 of 77.84% and an mAP of 77.61%. This suggests that the incorporation of all three loss components further augments the model's capability for accurate person reID.

4.4.5. Generalization with different backbones

We investigate CMOT adaptability by employing different feature extractor backbones and applying various fusion techniques for cross-modality fusion on the BIWI RGBD-ID. The backbones considered include Res2Net [97] DenseNet-201 [98] Inception-v3 [99] EfficientNet-B4 [98] and ResNet-50 [75]. The classifiers of each stream in the backbones are removed, and the modal features are flattened and passed through the fusion transformer, following the process depicted in Fig. 2. Table 10 presents the results of the cross-modality fusion transformer with different backbone networks, considering the number of parameters (M), rank-1, and mAP scores. The ResNet backbone achieves the best reID performance, with the CMOT fusion technique achieving a rank-1 of 74.88% and an mAP score of 72.98%. This performance surpasses other fusion techniques such as CM-Concat, ($CMOT_{SAMR}$), and ($CMOT_{CAMI}$). Similarly, when using the DenseNet-201 backbone, the CMOT fusion technique outperforms other techniques, achieving a rank-1 of 75.68% and a mAP score of 73.54%. In the case of the Inception-v3 backbone, implementing the CMOT fusion technique

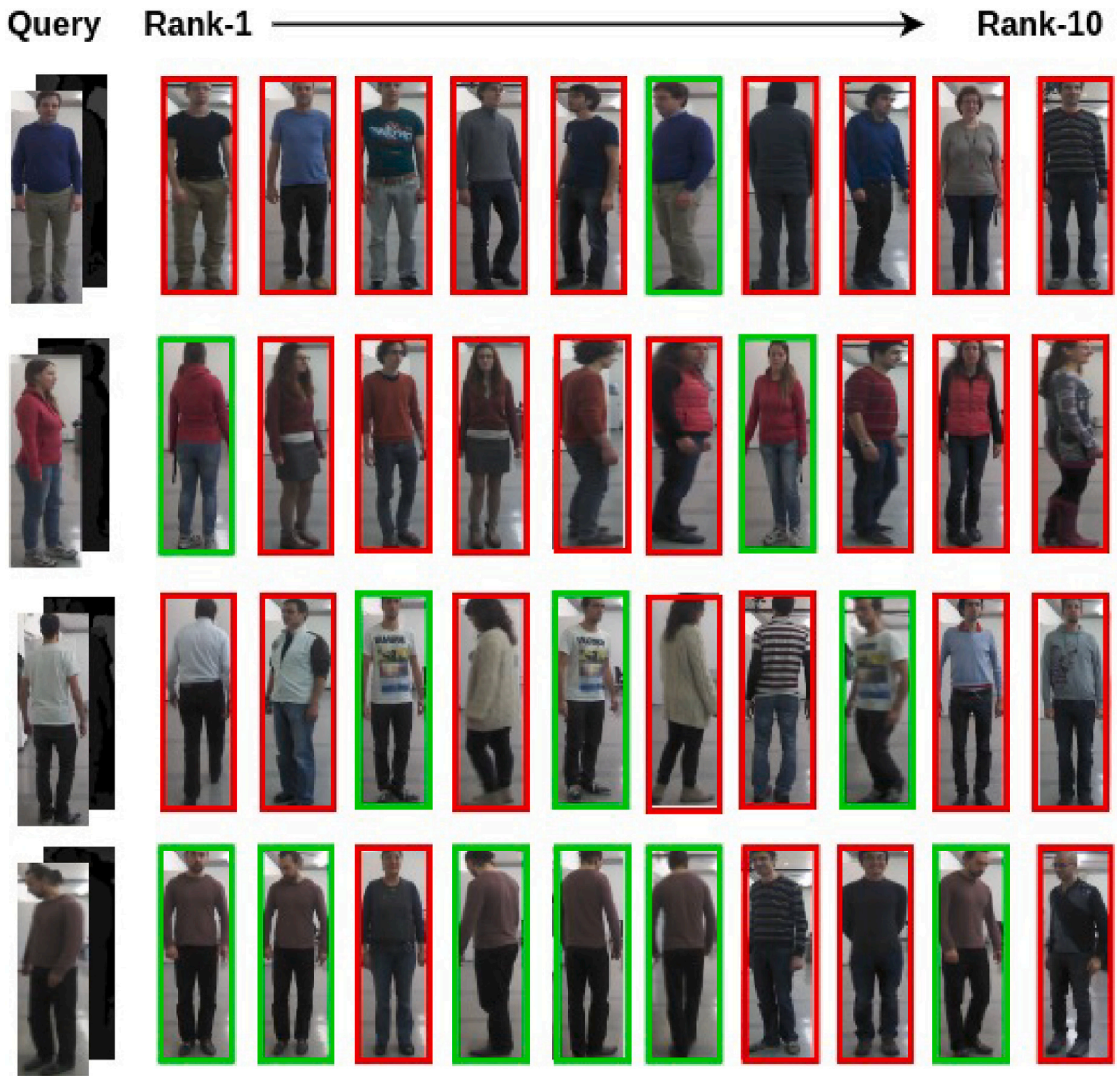


Fig. 8. The rank-10 retrieval results are obtained by our CMOT on the BIWI RGBD-ID dataset. The green and red bounding boxes refer to the correct and wrong matchings, respectively.

results in a rank-1 of 76.81% and a mAP score of 75.41%, showcasing better performance compared to alternative fusion techniques like CM-Concat, ($CMOT_{SAMR}$), and ($CMOT_{CAMI}$). Furthermore, with the EfficientNet-B4 backbone, the CMOT fusion technique demonstrates superior performance, achieving a rank-1 of 76.27% and a mAP score of 76.71%, surpassing other fusion techniques. Lastly, when employing the ResNet-50 backbone, the CMOT fusion technique attains a rank-1 of 77.84% and mAP score of 77.61%, outperforming other fusion techniques such as CM-Concat, ($CMOT_{SAMR}$), and ($CMOT_{CAMI}$). In conclusion, the comparison of various feature extraction backbones and fusion techniques in the CMOT model reveals that the CMOT fusion technique consistently performs better than other techniques across different backbones. This demonstrates the effectiveness of the CMOT fusion technique in handling RGB-D person re-identification tasks and achieving improved performance when compared to other fusion methods.

4.4.6. Qualitative analysis

We show the sample retrieval results of our proposed CMOT on the BIWI RGBD-ID dataset in Fig. 8. Gallery instances with the green bounding boxes predicted the same identity as the query input and red bounding boxes when the identities are different from the query. The outcomes corresponding to the first and second query instances illustrate that the model can predict wrong identities because of similar chest appearance, similar body structure, poster, or similar lower appearance. Additionally, the outcomes corresponding to the third and fourth query instances demonstrate that despite their similar appearance, CMOT is still able to identify them correctly which shows that CMOT can avoid wrong matching with similar appearances of query instances and retrieve positive instances. Overall, our CMOT correctly predicts the true identity at rank-10 effectively.

4.4.7. Modality-level attention map visualization

The visualization of the RGB, depth, and fused features is shown in Fig. 9 to explore the modality-level attention map in our CMOT. Based on the illustration in Fig. 9, it can be observed that our cross-modality fusion transformer predominantly concentrates on the torso regions, such as chest, belly, and pelvis, across both image modalities to extract the discriminative cross-modality features, while effectively ignoring the background regions. Furthermore, observation also reveals that activation feature maps of all modalities are analogous in nature. RGB mainly focuses on the upper body of the person, while upper

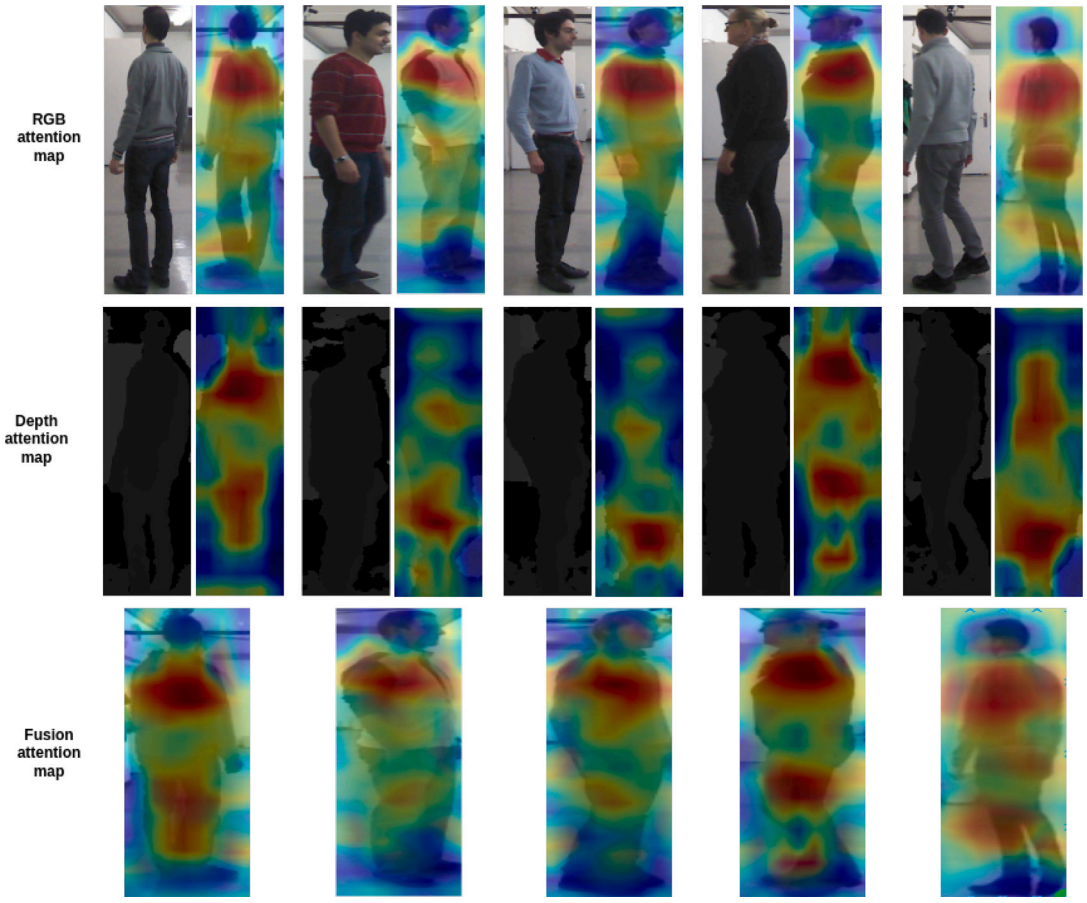


Fig. 9. The visual representations of attention maps are depicted for RGB, depth, and cross-modality fusion activations extracted after the fourth transformer layer of our CMOT. In the above row, RGB attention maps are shown where the left image in each pair shows the RGB input image and the right image is the activation map visualization. The middle row shows the depth modality activation map, while the last row is the fused modality activation map visualization. Within these feature visualization images, the intensity of brightness at a location is indicative of the strength of its activation, with brighter locales denoting stronger activations.

Table 10
Comparison of different fusion techniques with different backbone networks.

Backbone	Params (M)	Fusion technique	Rank-1	mAP
Res2Net [98]	25.5	CM-Concat	70.14	70.79
		$CMOT_{SAMR}$	70.81	70.00
		$CMOT_{CAMI}$	74.40	63.17
		CMOT	74.88	72.98
DenseNet-201 [99]	20	CM-Concat	71.54	70.07
		$CMOT_{SAMR}$	71.74	71.70
		$CMOT_{CAMI}$	74.42	63.38
		CMOT	75.68	73.54
Inception-v3 [100]	23.9	CM-Concat	72.61	71.11
		$CMOT_{SAMR}$	73.84	74.00
		$CMOT_{CAMI}$	75.08	64.84
		CMOT	76.81	75.41
EfficientNet-B4 [101]	19	CM-Concat	72.01	71.24
		$CMOT_{SAMR}$	74.64	74.41
		$CMOT_{CAMI}$	75.52	64.04
		CMOT	76.27	76.71
ResNet-50 [76]	25.6	CM-Concat	72.08	71.81
		$CMOT_{SAMR}$	75.21	75.50
		$CMOT_{CAMI}$	76.40	77.02
		CMOT	77.84	77.61

and lower body parts are the main focus in depth-modality activation maps. The fusion transformer is designed to pay attention to the joint discriminative features of RGB and depth modality and it is doing this efficiently as shown in the fusion attention maps. The regions accentuated by the features in fused modality are synonymous with the upper part, while lower body features are also involved due to depth modality. This is because the RGB modality is superior in offering appearance-based discriminative features.

5. Conclusion

This article proposes a novel framework called CMOT for RGB-D reID with RGB-D frame sequence. CMOT consists of dual-stream CNNs to extract features from each modality from the generated hypothesis and a fusion transformer for combining the cross-modality information, aiming to fully exploit the visual features from the RGB modality and corresponding depth information from the depth modality. By employing joint training with dual-stream CNNs, our approach maximizes the utilization of complementary information across modalities, resulting in the acquisition of a discriminative multimodal representation. The effectiveness of CMOT is achieved through the integration of a self-attention-driven SAMR and a cross-attention-driven CAMI. Additionally, the proposed BEFFB is also incorporated to enhance the reID performance while maintaining low computational overhead. The experimental results validate the importance and effectiveness of each component within CMOT. The findings demonstrate that CMOT exhibits versatility and scalability when applied with different CNNs. The performance evaluation showcases the significant contributions of SAMR, CAMI, and BEFFB in improving the reID accuracy and feature learning capabilities of CMOT. Furthermore, the results highlight the potential of CMOT for handling RGB-D person reID tasks across different datasets and modalities. However, future studies should focus on addressing the challenges associated with cross-modality learning

on more complex and small datasets and exploring the possibility of developing a reID framework based on the pure vision transformer that can process video without any external CNN for RGB-D reID involving more intricate scenarios and conditions for a person's reID. These complexities can include variations in lighting, backgrounds, poses, occlusions, and crowded scenes. Such challenges make the reID task more demanding and require advanced techniques to handle, while small datasets with fewer instances to learn from can challenge model generalization and robustness due to the scarcity of examples for capturing underlying patterns.

CRediT authorship contribution statement

Hamza Mukhtar: Writing – original draft, Methodology, Investigation, Conceptualization. **Muhammad Usman Ghani Khan:** Writing – review & editing, Validation, Supervision, Project administration, Data curation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Leonid Keselman, John Iselin Woodfill, Anders Grunnet-Jepsen, Achintya Bhowmik, Intel realsense stereoscopic depth cameras, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 1–10.

[2] Pushpajit Khaire, Praveen Kumar, A semi-supervised deep learning based video anomaly detection framework using RGB-D for surveillance of real-world critical environments, Forensic Sci. Int.: Digit. Investig. 40 (2022) 301346.

[3] Zhen Liu, Qin Cheng, Chengqun Song, Jun Cheng, Cross-scale cascade transformer for multimodal human action recognition, Pattern Recognit. Lett. 168 (2023) 17–23.

[4] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, Ben Upcroft, Simple online and realtime tracking, in: 2016 IEEE International Conference on Image Processing (ICIP), IEEE, 2016, pp. 3464–3468.

[5] Shujuan Wang, Bochun Huang, Huafeng Li, Guanqiu Qi, Dapeng Tao, Zhengtao Yu, Key point-aware occlusion suppression and semantic alignment for occluded person re-identification, Inform. Sci. 606 (2022) 669–687.

[6] Serhan Coşar, Nicola Bellotto, Human Re-identification with a robot thermal camera using entropy-based sampling, J. Intell. Robot. Syst. 98 (2020) 85–102.

[7] Jungong Han, Ling Shao, Dong Xu, Jamie Shotton, Enhanced computer vision with microsoft kinect sensor: A review, IEEE Trans. Cybern. 43 (5) (2013) 1318–1334.

[8] Dilin Liu, Hongxun Yao, Artistic image synthesis with tag-guided correlation matching, Multimedia Tools Appl. (2023) 1–12.

[9] Marina Paolanti, Rocco Pietrini, Adriano Mancini, Emanuele Frontoni, Primo Zingaretti, Deep understanding of shopper behaviours and interactions using RGB-D vision, Mach. Vis. Appl. 31 (2020) 1–21.

[10] Yongguo Ling, Zhun Zhong, Zhiming Luo, Paolo Rota, Shaozi Li, Nicu Sebe, Class-aware modality mix and center-guided metric learning for visible-thermal person re-identification, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 889–897.

[11] Md Kamal Uddin, Antony Lam, Hisato Fukuda, Yoshinori Kobayashi, Yoshinori Kuno, Fusion in dissimilarity space for RGB-D person re-identification, Array 12 (2021) 100089.

[12] Guowen Zhang, Pingping Zhang, Jinqing Qi, Huchuan Lu, Hat: Hierarchical aggregation transformers for person re-identification, in: Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 516–525.

[13] Hao Wang, Yiwen Sun, Xiaojun Bi, Structural redundancy reduction based efficient training for lightweight person re-identification, Inform. Sci. 637 (2023) 118962.

[14] Jingjing Wu, Jianguo Jiang, Meibin Qi, Cuiqun Chen, Jingjing Zhang, An end-to-end heterogeneous restraint network for RGB-D cross-modal person re-identification, ACM Trans. Multimed. Comput. Commun. Appl. (TOMM) 18 (4) (2022) 1–22.

[15] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, Shengjin Wang, Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline), in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 480–496.

[16] Chengmei Han, Bo Jiang, Jin Tang, Multi-granularity cross attention network for person re-identification, Multimedia Tools Appl. 82 (10) (2023) 14755–14773.

[17] Zhiqi Pang, Chunyu Wang, Junjie Wang, Lingling Zhao, Reliability modeling and contrastive learning for unsupervised person re-identification, Knowl.-Based Syst. (2023) 110263.

[18] Vladimir Somers, Christophe De Vleeschouwer, Alexandre Alahi, Body part-based representation learning for occluded person re-identification, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 1613–1623.

[19] Changxing Ding, Kan Wang, Pengfei Wang, Dacheng Tao, Multi-task learning with coarse priors for robust part-aware person re-identification, IEEE Trans. Pattern Anal. Mach. Intell. 44 (3) (2020) 1474–1488.

[20] Yaswanth Gavini, Arun Agarwal, B.M. Mehtre, Thermal to visual person re-identification using collaborative metric learning based on maximum margin matrix factorization, Pattern Recognit. 134 (2023) 109069.

[21] Amir Hadi, Makan Pourmasoumi, Ameneh Najafgholizadeh, Cain C.T. Clark, Ahmad Esmaillzadeh, The effect of apple cider vinegar on lipid profiles and glycemic parameters: a systematic review and meta-analysis of randomized clinical trials, BMC Complement. Med. Ther. 21 (1) (2021) 179.

[22] Xin Jin, Tianyu He, Kecheng Zheng, Zhiheng Yin, Xu Shen, Zhen Huang, Ruoyu Feng, Jianqiang Huang, Zhibo Chen, Xian-Sheng Hua, Cloth-changing person re-identification from a single image with gait prediction and regularization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 14278–14287.

[23] Chunsheng Hua, Xiaoheng Zhao, Wei Meng, Yingjie Pan, Deep person re-identification with the combination of physical biometric information and appearance features, in: Proceeding of 2021 International Conference on Wireless Communications, Networking and Applications, Springer, 2022, pp. 874–887.

[24] Dongjiang Cao, Ruofeng Liu, Hao Li, Shuai Wang, Wenchao Jiang, Chris Xiaoxuan Lu, Cross vision-rf gait re-identification with low-cost rgb-d cameras and mmwave radars, Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 6 (3) (2022) 1–25.

[25] Xiaoyu Tan, Weisheng Wang, Jinquan Wang, Hongli Lin, Semi-supervised person re-identification method based on voting algorithm, in: Proceedings of International Conference on Image, Vision and Intelligent Systems 2022 (ICIVIS 2022), Springer, 2023, pp. 722–732.

[26] Marina Paolanti, Roberto Pierdicca, Rocco Pietrini, Massimo Martini, Emanuele Frontoni, SeSAME: Re-identification-based ambient intelligence system for museum environment, Pattern Recognit. Lett. 161 (2022) 17–23.

[27] Nan Pu, Wei Chen, Yu Liu, Erwin M. Bakker, Michael S. Lew, Dual gaussian-based variational subspace disentanglement for visible-infrared person re-identification, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 2149–2158.

[28] Liangliang Ren, Jiwen Lu, Jianjiang Feng, Jie Zhou, Uniform and variational deep learning for RGB-D object recognition and person re-identification, IEEE Trans. Image Process. 28 (10) (2019) 4970–4983.

[29] 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, 2020, arXiv preprint arXiv:2010.11929.

[30] Aske R. Lejbolle, Kamal Nasrollahi, Benjamin Krogh, Thomas B. Moeslund, Multimodal neural network for overhead person re-identification, in: 2017 International Conference of the Biometrics Special Interest Group (BIOSIG), IEEE, 2017, pp. 1–5.

[31] Xing Fan, Wei Jiang, Hao Luo, Weijie Mao, Modality-transfer generative adversarial network and dual-level unified latent representation for visible thermal Person re-identification, Vis. Comput. (2022) 1–16.

[32] Qizao Wang, Xuelin Qian, Yanwei Fu, Xiangyang Xue, Co-attention aligned mutual cross-attention for cloth-changing person re-identification, in: Proceedings of the Asian Conference on Computer Vision, 2022, pp. 2270–2288.

[33] Haocong Rao, Chunyan Miao, SimMC: Simple masked contrastive learning of skeleton representations for unsupervised person re-identification, 2022, arXiv preprint arXiv:2204.09826.

[34] Junhui Yin, Zhanyu Ma, Jiyang Xie, Shibo Nie, Kongming Liang, Jun Guo, Dual-granularity feature alignment for cross-modality person re-identification, Neurocomputing 511 (2022) 78–90.

[35] Can Zhang, Hong Liu, Wei Guo, Mang Ye, Multi-scale cascading network with compact feature learning for RGB-infrared person re-identification, in: 2020 25th International Conference on Pattern Recognition (ICPR), IEEE, 2021, pp. 8679–8686.

[36] Xin Xu, Xin Yuan, Zheng Wang, Kai Zhang, Ruimin Hu, Rank-in-rank loss for person re-identification, ACM Trans. Multimed. Comput. Commun. Appl. (TOMM) 18 (2s) (2022) 1–21.

[37] Maged Shoman, Armstrong Aboah, Alex Morehead, Ye Duan, Abdulateef Daud, Yaw Adu-Gyamfi, A region-based deep learning approach to automated retail checkout, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 3210–3215.

[38] Guoqing Zhang, Weisi Lin, Arun kumar Chandran, Xuan Jing, Complementary networks for person re-identification, Inform. Sci. 633 (2023) 70–84.

[39] Albert Haque, Alexandre Alahi, Li Fei-Fei, Recurrent attention models for depth-based person identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1229–1238.

[40] Haocong Rao, Chunyan Miao, TranSG: Transformer-based skeleton graph prototype contrastive learning with structure-trajectory prompted reconstruction for person re-identification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 22118–22128.
[41] Ancong Wu, Wei-Shi Zheng, Jian-Huang Lai, Robust depth-based person re-identification, IEEE Trans. Image Process. 26 (6) (2017) 2588–2603.
[42] Wei-Yu Lee, Ljubomir Jovanov, Wilfried Philips, Cross-modality attention and multimodal fusion transformer for pedestrian detection, in: Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part V, Springer, 2023, pp. 608–623.
[43] Kuan Zhu, Haiyun Guo, Shiliang Zhang, Yaowei Wang, Jing Liu, Jinqiao Wang, Ming Tang, Aaformer: Auto-aligned transformer for person re-identification, IEEE Trans. Neural Netw. Learn. Syst. (2023).
[44] Zhicheng Sun, Yadong Mu, Patch-based knowledge distillation for lifelong person re-identification, in: Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 696–707.
[45] Jinze Huang, Xiaohan Yu, Dong An, Yaoguang Wei, Xiao Bai, Jin Zheng, Chen Wang, Jun Zhou, Learning consistent region features for lifelong person re-identification, Pattern Recognit. 144 (2023) 109837.
[46] Matteo Munaro, Andrea Fossati, Alberto Basso, Emanuele Menegatti, Luc Van Gool, One-shot person re-identification with a consumer depth camera, in: Person Re-Identification, Springer, 2014, pp. 161–181.
[47] Hong Liu, Liang Hu, Liqian Ma, Online RGB-D person re-identification based on metric model update, CAAI Trans. Intell. Technol. 2 (1) (2017) 48–55.
[48] Shutao Bai, Bingpeng Ma, Hong Chang, Rui Huang, Xilin Chen, Salient-to-broad transition for video person re-identification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 7339–7348.
[49] Jing Zhao, Long Lan, Da Huang, Jing Ren, Wenjing Yang, Heterogeneous pseudo-supervised learning for few-shot person re-identification, Neural Netw. 154 (2022) 521–537.
[50] Xinyu Zhang, Dongdong Li, Zhigang Wang, Jian Wang, Errui Ding, Javen Qinfeng Shi, Zhaoxiang Zhang, Jingdong Wang, Implicit sample extension for unsupervised person re-identification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 7369–7378.
[51] Yunhua Lu, Mingzi Jiang, Zhi Liu, Xinyu Mu, Dual-branch adaptive attention transformer for occluded person re-identification, Image Vis. Comput. 131 (2023) 104633.
[52] Weihua Chen, Xiaotang Chen, Jianguo Zhang, Kaiqi Huang, Beyond triplet loss: a deep quadruplet network for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 403–412.
[53] Honghu Pan, Yang Bai, Zhenyu He, Chunkai Zhang, AAGCN: Adjacency-aware graph convolutional network for person re-identification, Knowl.-Based Syst. 236 (2022) 107300.
[54] Yuqiao Xian, Jinrui Yang, Fufu Yu, Jun Zhang, Xing Sun, Graph-based self-learning for robust person re-identification, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 4789–4798.
[55] Yue Zhang, Fanghui Zhang, Yi Jin, Yigang Cen, Viacheslav Voronin, Shaohua Wan, Local correlation ensemble with GCN based on attention features for cross-domain person Re-ID, ACM Trans. Multimed. Comput. Commun. Appl. 19 (2) (2023) 1–22.
[56] Nikolaos Karianakis, Zicheng Liu, Yinpeng Chen, Stefano Soatto, Reinforced temporal attention and split-rate transfer for depth-based person re-identification, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 715–733.
[57] Zeynab Imani, Hadi Soltanizadeh, Ali A. Orouji, Tensor-based sparse canonical correlation analysis via low rank matrix approximation for RGB-D long-term person re-identification, Multimedia Tools Appl. 79 (2020) 11787–11811.
[58] Zeynab Imani, Hadi Soltanizadeh, Person reidentification using local pattern descriptors and anthropometric measures from videos of kinect sensor, IEEE Sens. J. 16 (16) (2016) 6227–6238.
[59] Komal Soni, Debi Prosad Dogra, Arif Ahmed Sekh, Samarjit Kar, Heeseung Choi, Ig-Jae Kim, Person re-identification in indoor videos by information fusion using Graph Convolutional Networks, Expert Syst. Appl. 210 (2022) 118363.
[60] Zi Wang, Chenglong Li, Aihua Zheng, Ran He, Jin Tang, Interact, embed, and enlarge: boosting modality-specific representations for multi-modal person re-identification, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36, 2022, pp. 2633–2641.
[61] Xin Hao, Sanyuan Zhao, Mang Ye, Jianbing Shen, Cross-modality person re-identification via modality confusion and center aggregation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 16403–16412.
[62] Seokeon Choi, Sumin Lee, Youngeun Kim, Taekyung Kim, 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, 2020, pp. 10257–10266.
[63] Xuezhi Xiang, Ning Lv, Zeting Yu, Mingliang Zhai, Abdulmotaleb El Saddik, Cross-modality person re-identification based on dual-path multi-branch network, IEEE Sens. J. 19 (23) (2019) 11706–11713.
[64] Yuanxin Zhu, Zhao Yang, Li Wang, Sai Zhao, Xiao Hu, Dapeng Tao, Hetero-center loss for cross-modality person re-identification, Neurocomputing 386 (2020) 97–109.
[65] Ammarah Farooq, Muhammad Awais, Josef Kittler, Syed Safwan Khalid, AXM-Net: Implicit cross-modal feature alignment for person re-identification, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36, 2022, pp. 4477–4485.
[66] Liangliang Ren, Jiwen Lu, Jianjiang Feng, Jie Zhou, Multi-modal uniform deep learning for RGB-D person re-identification, Pattern Recognit. 72 (2017) 446–457.
[67] Yan Huang, Qiang Wu, JingSong Xu, Yi Zhong, ZhaoXiang Zhang, Clothing status awareness for long-term person re-identification, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 11895–11904.
[68] Wenjie Zhen, Lin Yang, Mei-Po Kwan, Zejun Zuo, Bo Wan, Shunping Zhou, Shengwen Li, Yaqin Ye, Haoyue Qian, Xiaofang Pan, Capturing what human eyes perceive: A visual hierarchy generation approach to emulating saliency-based visual attention for grid-like urban street networks, Comput. Environ. Urban Syst. 80 (2020) 101454.
[69] Matteo Rizzo, Cristina Conati, Daesik Jang, Hui Hu, Evaluating the faithfulness of saliency-based explanations for deep learning models for temporal colour constancy, 2022, arXiv preprint arXiv:2211.07982.
[70] Shakeeb Murtaza, Soufiane Belharbi, Marco Pedersoli, Aydin Sarraf, Eric Granger, Discriminative sampling of proposals in self-supervised transformers for weakly supervised object localization, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 155–165.
[71] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).
[72] Dmitrii Marin, Jen-Hao Rick Chang, Anurag Ranjan, Anish Prabhu, Mohammad Rastegari, Oncel Tuzel, Token pooling in vision transformers, 2021, arXiv preprint arXiv:2110.03860.
[73] Jiachen Li, Menglin Wang, Xiaojin Gong, Transformer based multi-grained features for unsupervised person re-identification, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 42–50.
[74] Shenqi Lai, Zhenhua Chai, Xiaolin Wei, Transformer meets part model: Adaptive part division for person re-identification, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 4150–4157.
[75] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, Xi Zhou, Learning discriminative features with multiple granularities for person re-identification, in: Proceedings of the 26th ACM International Conference on Multimedia, 2018, pp. 274–282.
[76] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[77] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, Wei Jiang, Transreid: Transformer-based object re-identification, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 15013–15022.
[78] Zehua Chai, Yongguo Ling, Zhiming Luo, Dazhen Lin, Min Jiang, Shaozi Li, Dual-stream transformer with distribution alignment for visible-infrared person re-identification, IEEE Trans. Circuits Syst. Video Technol. (2023).
[79] Haoyan Ma, Xiang Li, Xia Yuan, Chunxia Zhao, Denseformer: A dense transformer framework for person re-identification, IET Comput. Vis. 17 (5) (2023) 527–536.
[80] Wanyin Wu, Dapeng Tao, Hao Li, Zhao Yang, Jun Cheng, Deep features for person re-identification on metric learning, Pattern Recognit. 110 (2021) 107424.
[81] Christopher Neff, Armin Danesh Pazho, Hamed Tabkhi, Towards real-time online unsupervised domain adaptation for real-world person re-identification, 2023.
[82] Xinyu Yang, Tilo Burghardt, Majid Mirmehdi, Dynamic curriculum learning for great ape detection in the wild, Int. J. Comput. Vis. (2023) 1–19.
[83] Yunhao Du, Zhicheng Zhao, Yang Song, Yanyun Zhao, Fei Su, Tao Gong, Hongying Meng, Strongsort: Make deepsort great again, IEEE Trans. Multimed. (2023).
[84] Nicolai Wojke, Alex Bewley, Dietrich Paulus, Simple online and realtime tracking with a deep association metric, in: 2017 IEEE International Conference on Image Processing (ICIP), IEEE, 2017, pp. 3645–3649.
[85] Leqi Shen, Tao He, Yuchen Guo, Guiguang Ding, X-ReID: Cross-instance transformer for identity-level person re-identification, 2023, arXiv preprint arXiv:2302.02075.
[86] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, Qi Tian, Deep modular co-attention networks for visual question answering, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 6281–6290.
[87] Guoqing Zhang, Junchuan Yang, Yuhui Zheng, Ye Wang, Yi Wu, Shengyong Chen, Hybrid-attention guided network with multiple resolution features for person re-identification, Inform. Sci. 578 (2021) 525–538.
[88] Zihan Guo, Dezhi Han, Sparse co-attention visual question answering networks based on thresholds, Appl. Intell. 53 (1) (2023) 586–600.

[89] Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, Tieyan Liu, On layer normalization in the transformer architecture, in: International Conference on Machine Learning, PMLR, 2020, pp. 10524–10533.

[90] Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong, Jianhuang Lai, RGB-infrared cross-modality person re-identification, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 5380–5389.

[91] Haocong Rao, Xiping Hu, Jun Cheng, Bin Hu, SM-SGE: A self-supervised multi-scale skeleton graph encoding framework for person re-identification, in: Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 1812–1820.

[92] Frank M. Hafner, Amran Bhuyian, Julian F.P. Kooij, Eric Granger, Cross-modal distillation for RGB-depth person re-identification, Comput. Vis. Image Underst. 216 (2022) 103352.

[93] Haocong Rao, Cyril Leung, Chunyan Miao, Hierarchical skeleton meta-prototype contrastive learning with hard skeleton mining for unsupervised person re-identification, 2023, arXiv preprint arXiv:2307.12917.

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

[95] Aske Rasch Lejbølle, Kamal Nasrollahi, Benjamin Krogh, Thomas B. Moeslund, Person re-identification using spatial and layer-wise attention, IEEE Trans. Inf. Forensics Secur. 15 (2019) 1216–1231.

[96] Ankur Bapna, Mia Xu Chen, Orhan Firat, Yuan Cao, Yonghui Wu, Training deeper neural machine translation models with transparent attention, 2018, arXiv preprint arXiv:1808.07561.

[97] Guan'an Wang, Tianzhu Zhang, Jian Cheng, Si Liu, Yang Yang, 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, 2019, pp. 3623–3632.

[98] Shang-Hua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, Philip Torr, Res2net: A new multi-scale backbone architecture, IEEE Trans. Pattern Anal. Mach. Intell. 43 (2) (2019) 652–662.

[99] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, Kilian Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4700–4708.

[100] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.

[101] Mingxing Tan, Quoc Le, Efficientnet: Rethinking model scaling for convolutional neural networks, in: International Conference on Machine Learning, PMLR, 2019, pp. 6105–6114.