# Transformer-Based Person Re-Identification: A Comprehensive Review

Prodip Kumar Sarker, Qingjie Zhao, and Md. Kamal Uddin

*Abstract*—**In the evolving landscape of surveillance and security applications, the task of person re-identification(re-ID) has significant importance, but also presents notable difficulties. This task entails the process of accurately matching and identifying persons across several camera views that do not overlap with one another. This is of utmost importance to video surveillance, public safety, and person-tracking applications. However, vision-related difficulties, such as variations in appearance, occlusions, viewpoint changes, cloth changes, scalability, limited robustness to environmental factors, and lack of generalizations, still hinder the development of reliable person re-ID methods. There are few approaches have been developed based on these difficulties relied on traditional deep-learning techniques. Nevertheless, recent advancements of transformer-based methods, have gained widespread adoption in various domains owing to their unique architectural properties. Recently, few transformer-based person re-ID methods have developed based on these difficulties and achieved good results. To develop reliable solutions for person re-ID, a comprehensive analysis of transformer-based methods is necessary. However, there are few studies that consider transformer-based techniques for further investigation. This review proposes recent literature on transformer-based approaches, examining their effectiveness, advantages, and potential challenges. This review is the first of its kind to provide insights into the revolutionary transformer-based methodologies used to tackle many obstacles in person re-ID, providing a forward-thinking outlook on current research and potentially guiding the creation of viable applications in real-world scenarios. The main objective is to provide a useful resource for academics and practitioners engaged in person re-ID.**

*Index Terms*—**Person re-identification, transformer, vision transformer, challenges and evaluation matrices.**

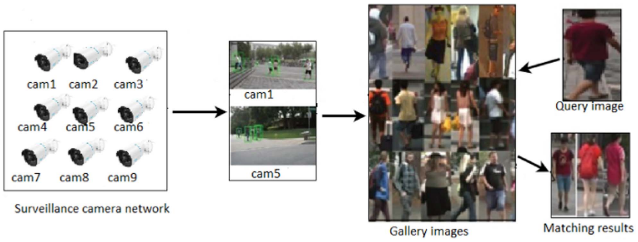

Fig. 1. Overall framework of the process of person re-identification.

## I. Introduction

A PERSON re-ID is related to the process of identifying and monitoring persons across several non-overlapping camera networks situated in outdoor and indoor environments. Fig. 1 illustrates the general framework of person re-ID techniques. Person re-ID has attracted significant attention from both academia and business in the context of surveillance systems [1], public safety applications, and urban smart city deployments, where identifying people across several camera views and at various times is crucial. While surveillance and security are the areas where person re-ID is most immediately utilized, other fields such as retail (understanding client mobility), event management (observing VIPs or specific persons), and smart transportation systems may also benefit from it. Gheissari et al. [2] initially suggested idea of person re-ID in 2006. Later, lot of researchers developed innovative algorithms, models, and datasets to effectively tackle re-ID challenges while improving the overall performance. The aforementioned developments are currently in progress in the areas of feature learning, DL architectures, metric learning, domain adaptation, and multimodal fusion addresses in the context of person re-ID. Person re-ID [3], [4], [5], [6], [7], [8], [9], [10], [11], [12] involves comparing and matching photos of the same pedestrian captured by many cameras with varying focal lengths or color filters. Traditionally, person re-ID has relied on manual annotations such as bounding boxes or manual feature extraction, which are time-consuming and not scalable for large-scale deployments. Additionally, conventional methods have struggled to handle appearance changes caused by variations in pose, illumination, and camera viewpoints, leading to limited accuracy and robustness, and this issue remains unresolved in terms of its practical use in real-life situations. A transformer first gathers local and global information to extract robust characteristics. Recently, person re-ID has seen significant advancements in terms of performance and implementation in practice, generally owing to the progress made in methods based on transformer [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23]. Transformer based approaches are beneficial for person re-ID tasks involving diverse modalities, ultimately leading to the achievement of visible-infrared person VI-reID.

Despite advancements in person re-ID, numerous notable challenges remain. These challenges include effectively managing extensive datasets, addressing the difficulties caused by

Manuscript received 22 November 2023; accepted 28 December 2023. Date of publication 8 January 2024; date of current version 25 September 2024. *(Corresponding author: Qingjie Zhao.)*

Prodip Kumar Sarker and Qingjie Zhao are with the School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China (e-mail: prodip@bit.edu.cn; zhaoqj@bit.edu.cn).

Md. Kamal Uddin is with the Department of Computer Science and Telecommunication Engineering, Noakhali Science and Technology University, Noakhali 3814, Bangladesh (e-mail: kamaluddin@nstu.edu.bd).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIV.2024.3350669.

Digital Object Identifier 10.1109/TIV.2024.3350669

TABLE I
COMPARATIVE ANALYSIS OF SEVERAL PERSON RE-ID SURVEYS

| Survey | References | Contribution |
|---|---|---|
| Wang et al. [24] | CAAI,18 | I. The authors reviewed Hybrid, CNN, and GAN-based re-ID architectures.<br>II. They suggested directions for further investigation.<br>III. Number of articles used: 62. |
| Wu et al. [25] | NC, 19 | I. Examine six DL methods: identification, verification, distance metric, component, video,and data augmentation models.<br>II. Summarized futuristic results on many typical re-ID datasets.<br>III. Number of articles used: 154. |
| Almasawa et al. [26] | Access, 19 | I. Reviewed image-based, video-based, or image-to-video hybrid DL methods.<br>II. Using advanced algorithms to evaluate and contrast the top-ranking accuracy results obtained using standard datasets.<br>III. Number of articles 101. |
| Leng et al. [27] | CSVT, 19 | I. Examined closed-world and open-world differences.<br>II. Discussed open-world re-ID advancements and limitations.<br>III. 111 number of articles used in this review |
| Mathur et al. [28] | ICETCE, 20 | I. This review presents some challenges that DL methods.<br>II overcomes. To handle diverse viewpoints, dazedness, and resolutions in the image.<br>III. Number of articles used: 85. |
| Islam et al. [29] | IVC, 20 | I. Discussed deep metric learning with novel loss functions and feature representation learning.<br>II. Number of articles used 86. |
| Wang et al. [30] | AIC, 21 | I. Examined cross-domain person re-ID approaches.<br>II. Comparing these approaches' performance on publicly available datasets.<br>III. The number of articles used was 35. |
| Yang et al. [31] | CDS, 21 | I. Examined modern unsupervised person re-ID.<br>II. Compared several techniques in-depth based on the appropriate category.<br>III.There are 34 articles used in this review. |
| Ye et al. [3] | PAMI, 21 | I. The issues of re-ID viewpoints were considered when analyzing both open and closed-world situations.<br>II. A total of 255 articles were used in this review. |
| Peng et al. [32] | arXiv, 22 | I. Summarizing and evaluating a person's performance occluded to the mainstream based on their requirements, re-ID methods provided to academics, and businesses.<br>II. There are 106 articles reviewed in this paper |
| Huang et al. [33] | IF, 22 | I. Presented the history of VI re-ID and several state-of-the-art technique comparison studies on benchmark datasets.<br>III. Articles used in this survey were 125. |
| Ming et al. [34] | IVC, 22 | I. Examined four deep learning-based approaches for re-identifying people: generative adversarial, sequence feature learning, deep metric, and local feature.<br>II. Examined how well various approaches performed on the provided datasets.<br>III. Article reviewed in this paper: 205. |
| Zahra et al. [35] | PR, 23 | I. To resolve the numerous obstacles associated with person re-ID.<br>II. Analyzed supervised, unsupervised and semi-supervised techniques on image, video and synthetic datasets.<br>III. There are a total of 281 articles reviewed in this survey. |

variations in pose and appearance, reducing the adverse impacts of occlusions and background clutter, and attaining robustness across diverse camera views and environmental conditions. Resolving these difficulties using transformer based methods is of utmost importance to facilitate the development of person re-ID systems that are both precise and dependable, consequently facilitating their successful application in realistic scenarios. In recent years, a few investigations on person re-ID have been published [3], [24], [25], [26], [27], [28], [29], [30], [31], [32], [33], [34], [35]. Each concentrates on a unique facet of the re-ID issue. All the aforementioned reviews carefully examined the benefits and limitations of several methods under different setting considerations. Furthermore, the authors presented valuable suggestions for future research. Recently transformer based methods served as the basis for numerous advanced algorithms in the field of person re-ID to address person re-ID challenges. However, there is a lack of comprehensive reviews on the impact of transformer-based methodologies on performance outcomes and the significance of datasets in addressing various vision-related difficulties. For this perspective, we encouraged to write a person re-ID review paper based on transformer based methods. Table I illustrates how our survey differs from those of other studies in this category.The purpose of our study is to analyzed the current conventional DL methods and transformer-based approaches in detail, organized based on the particular challenges addressed. Furthermore, according to the progress reviewed, the challenges highlighted in Fig. 2, which have achieved state-of-the-art achievements in certain problems, are critically reviewed. The main contributions of our transformer-based person re-ID review are:

- We critically analyzed and investigate person re-ID based on transformer-based methods and its variants. We also discussed the advantages of transformer based models in person re-ID. As far as we know, nobody has yet summarized them.
- We explored transformer based methods to determine the obstacles. We compared the performance of transformer-based models with traditional methods on benchmark datasets
- The main challenges in the domain of human re-ID are summarized, while it is acknowledged that there exists a significant need for more investigation in this area. Additionally, we provide an overview of potential directions for future investigation in the field of human re-ID.

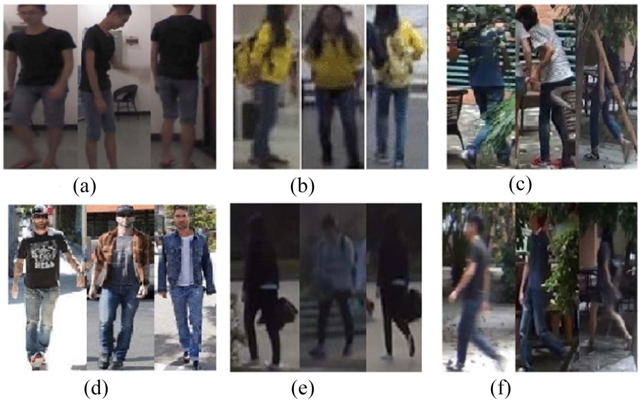

Fig. 2. Challenges of person re-ID problem owing to camera and illumination variations are shown in several images. (a) Variations in Appearance. (b) Viewpoint Changes. (c) Occlusions. (d) Clothing. (e) Limited Robustness to Environmental Factors. (f) Lack of Generalization.

The remaining part of our review is organized as follows: In Section II discusses the survey methodology in case of person re-ID. After that, details of the most widely used person re-ID datasets and evaluation metrics are provided in sections III. Sections IV and V represents the loss function and general structures of person re-ID. In Section VI comprehensive evaluation and experimental comparison methods on benchmark datasets are discussed. Finally, some challenges and future directions are discussed in Section VII, and the conclusion is made in the final section.

## II. Person Re-Id Based Survey Methodology

Re-identification of a person aims to accurately re-identify individuals in images acquired by different cameras or at various times. This involves extracting discriminative features from images or videos and comparing them to determine whether two images depict the same person. The survey methodology in person re-ID refers to collecting and analyzing articles to compare the performance of re-ID techniques. It is important to note that survey methodology may vary depending on the specific research objectives and state of the field. Researchers must ensure the rigor and reproducibility of their experiments and follow ethical data collection and usage guidelines.

### A. Selection of Research

PRISMA provides guidelines for reporting systematic reviews and meta-analyses, including defining a research question, developing an inclusion and exclusion criteria list, conducting a comprehensive literature search, removing duplicate records, and screening titles and abstracts against inclusion and exclusion criteria. Full-text articles were obtained, assessed for eligibility, resolved conflicts and discrepancies, documented the study selection process, extracted relevant data, and adhered to PRISMA guidelines to ensure quality and clarity. We followed the PRISMA guidelines [36] when selecting the studies for this meta-analysis [37], [38], [39]. Computer science review articles can still benefit from PRISMA's standard and systematic approach for publishing systematic literature reviews. PRISMA was initially devised for health research reporting. By adhering to the PRISMA guidelines, researchers can increase the clarity and rigor of the systematic review method they use as well as ensure the inclusion of essential elements such as the search approach, selection criteria, and quality assessment. This has the potential to strengthen the credibility of assessments and their findings, regardless of the subject or academic field. The process of selecting papers for this review is summarized in Fig. 3. We selected the manuscripts in two stages using the PRISMA guidelines for this evaluation. In the first round, all the items that did not satisfy the criteria were eliminated. In the second phase, relevant publications were identified using full-text reports. Once we had a manageable list of papers, we eliminated those that used datasets or performance metrics that were not typically accepted in the scientific literature.

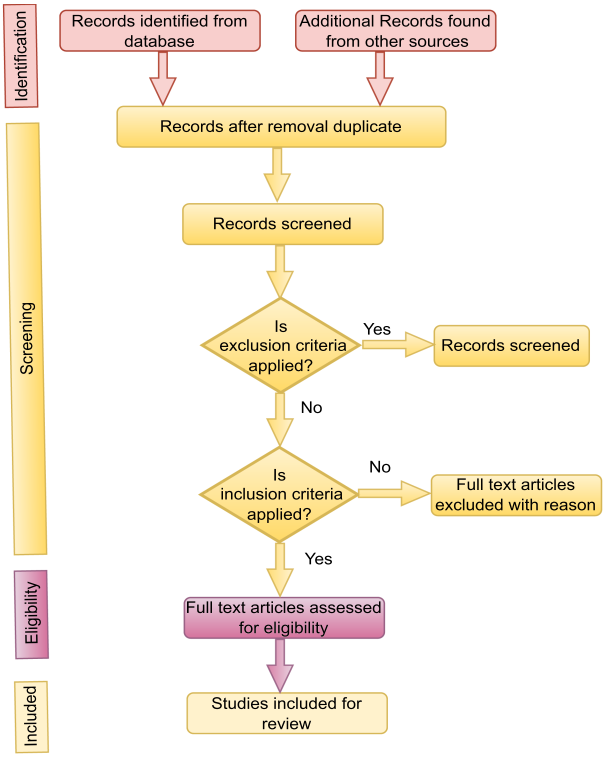
Identification
Records identified from database
Additional Records found from other sources
Records after removal duplicate
Screening
Records screened
Is exclusion criteria applied?
Yes
Records screened
No
Is inclusion criteria applied?
No
Full text articles excluded with reason
Yes
Eligibility
Full text articles assessed for eligibility
Included
Studies included for review

Fig. 3. Summary of articles selection process for transformer-based person re-ID review.

### B. Techniques for the Extraction of Data

Data extraction methods are techniques and processes used to collect data from various sources. The specific data extraction methods employed can vary depending on the nature of the data, sources from which the data are collected, and objectives of the study. A preliminary list of articles for this evaluation was compiled by thoroughly searching relevant databases such as PubMed, Embase, Springer, Scopus, Google Scholar, Elsevier, IEEE Xplore, and Web of Science to identify potential studies. Additionally, we searched the gray literature, conference proceedings, and contact experts in the field for additional

TABLE II
CRITERIA USED FOR THE EXCLUSION AND INCLUSION OF SELECTED ARTICLES

| Criteria for inclusion | Criteria for exclusion |
|---|---|
| Include articles that specifically address person re-ID as the primary research focus. | Exclude publications that are not about person re-ID or the research topic. |
| Include empirical research using experiments, assessments, or case studies. | Literature reviews, surveys, and systematic reviews are not research investigations. |
| Include papers that evaluate new or publicly accessible person re-ID datasets. | Exclude articles that only appear as conference abstracts without full-length publications. |
| Include articles on feature extraction, matching algorithms, and deep learning models for person re-id. | Avoid duplicate papers or studies from the same study group with comparable findings. |
| Quantitative analyses of person re-ID approaches employing rank-1 accuracy, mean Average Precision, or CMC curves are included. | Specify the language(s) for included or excluded articles based on the review's scope. |

studies. The top journals and conferences were considered for this review. The following terms were used to find pertinent articles:

- Person re-ID
- Supervised person re-ID
- Semi-supervised person re-ID
- Unsupervised person re-ID
- Conventional deep learning based person re-ID
- Pose variations
- Clothing person re-ID
- Occlusion-based person re-ID
- Transformer-based re-ID
- Vision transformer based re-ID
- End-to-end learning

The inclusion or exclusion of papers was determined as shown in Table II. We begin our decision-making process by using the headline. If an article's title did not pass muster, we examined the abstract and conclusion to determine if it was acceptable. To continue our investigation, we have immersed ourselves in extensive literature.

### C. The Compilation of Data

To make our study more valuable and to invite contributions from other academics, we prepared a data extraction sheet outlining the numerous important data pieces to be obtained from the papers. Around 25 data items were used to collect metadata related information from each article. Metadata include sources of article, different challenges, methodological particulars, reproducibility, code accessibility, performance metrics, and the datasets used. The results will be documented in a spreadsheet and made accessible to researchers who want to dig deeper or examine the topic from a new viewpoint based on the information provided.

### D. Results

By using the selection criteria, 450 articles were obtained from different sources. The database now contains 425 articles after eliminating unnecessary articles. More than 249 articles were rejected based on our inclusion and exclusion criteria, providing 176 participants for the evaluation. We have added relevant publications on persons' re-ID from places including AAAI, NIPS, IJCAI, ACM MM, ICLR, and others.

## III. DATASETS AND EVALUATION METRICS

When dealing with person re-ID, it is important to take into account the features of the particular datasets and choose evaluation metrics that meet the needs of the application. Person re-ID methods are usually assessed based on certain metrics and datasets.

### A. Person re-ID Datasets

Person re-ID datasets are designed to facilitate research and development in computer vision, specifically focusing on identifying and tracking individuals across different camera views or images. These datasets typically consist of images or videos captured from multiple cameras in various environments such as campuses, streets, airports, and shopping malls. In the context of transformer-based person re-ID, two different modality datasets are used: single-modality and cross-modality datasets.

Single-modality datasets help study the effectiveness of specific visual cues or sensors in person re-ID [40], [41], [42], [43], [44], [45], [46], [47], [48], [49], [50], [51], [52], [53], [54], [55], [56], [57]. They allow researchers to develop algorithms and models tailored to the characteristics and limitations of a particular modality. Cross-modality datasets allow researchers to explore the advantages and challenges of integrating multiple sensor modalities for person re-ID [58], [59], [60]. Combining different data sources enables these datasets to develop algorithms and models that leverage complementary information and improve performance in challenging scenarios, such as varying lighting conditions, pose variations, or occlusions. Table III and Table IV show the summarized information of different single-modality re-ID and visible-infrared cross-modality datasets.

### B. Performance Evaluation Metrics

The precision and efficacy of person re-ID algorithms were evaluated using a variety of performance evaluation metrics [3], [50], [61]. Some of these metrics are as follows:

TABLE III
SUMMARY OF DIFFERENT SINGLE MODALITY PERSON RE-ID DATASETS

| Datasets | Release time | Number of Identities | Number of Cameras | Number of images | Label | Crop size |
|---|---|---|---|---|---|---|
| VIPeR [40] | 2007 | 632 | 2 | 1264 | Hand | 128X48 |
| GRID [41] | 2009 | 1025 | 8 | 1275 | Hand | Vary |
| CAVIAR4ReID [42] | 2011 | 72 | 2 | 1220 | Hand | Vary |
| PRID2011 [41] | 2011 | 934 | 2 | 24541 | Hand | 128X64 |
| V47 [43] | 2011 | 47 | 2 | 752 | Hand | Vary |
| WARD [44] | 2012 | 70 | 3 | 4786 | Hand | 128X48 |
| CUHK01 [45] | 2012 | 971 | 2 | 3884 | Hand | 160X60 |
| CUHK02 [46] | 2013 | 1816 | 10(5 pairs) | 7264 | Hand | 160X60 |
| CUHK03 [47] | 2014 | 1467 | 10(5 pairs) | 13164 | Hand/DPM | Vary |
| RAiD [48] | 2014 | 43 | 4 | 6920 | Hand | 128X64 |
| iLIDS-VID [49] | 2014 | 300 | 2 | 42495 | Hand | Vary |
| Market1501 [50] | 2015 | 1501 | 6 | 32217 | Hand/DPM | 128X64 |
| PKU-Reid [51] | 2016 | 114 | 2 | 1824 | Hand | 128X64 |
| PRW [52] | 2016 | 932 | 6 | 34304 | Hand | vary |
| DukeMTMC-reID [53] | 2017 | 1812 | 8 | 36441 | Hand | Vary |
| Airport [54] | 2017 | 9651 | 6 | 39902 | ACF | 128X64 |
| MSMT17 [55] | 2018 | 4101 | 15 | 126441 | Faster RCNN | Vary |
| LPW [56] | 2018 | 2,731 | 3,4,4 | 592,438 | Detector+NN+Hand | - |
| PKU SketchRe-ID [57] | 2018 | 200 | 2 | 400 | Hand | - |

TABLE IV
SUMMARY OF THREE VISIBLE-INFRARED CROSS-MODALITY DATASETS

| Datasets | Release Time | Number of Identities | Number RGB Cameras | Number Infrared Cameras | Number of Images |
|---|---|---|---|---|---|
| SYSU-MM01 [58] | 2017 | 491 | 4 | 2 | 38271 |
| RegDB [59] | 2017 | 412 | 1 | 1 | 8240 |
| LLCM [60] | 2023 | 1064 | 9 | 9 | 46767 |

*1) Cumulative Matching Characteristics:* In cumulative matching characteristics(CMC), rank-r denotes how likely a match will be found among top-rank search results arranged. CMC [61] cannot effectively describe a model's discriminability for multiple shots, because it only counts the first-ranked match. In the single-shot scenario, the re-ID model method is given a query picture and orders all gallery photos from the most to the least similar. For this query picture, the top-r CMC accuracy $ACC_r$ is determined as follows:

$$ACC_r = \begin{cases} 1, & \text{if top } - r \text{ gallery images have query identity} \\ 0, & \text{Or else} \end{cases} \tag{1}$$

The overall score was calculated as follows:

$$CMC_r = \frac{\sum_{q=1}^{N_Q} ACC_k}{N_Q} \tag{2}$$

Where $r \in 1, 2, 3, \ldots\ldots\ldots, N_G$. $N_G$ and $N_Q$ denote the number of photographs in the gallery and question, respectively.

*2) Mean Average Precision:* Mean average precision (mAP) [50] is a complete statistic for evaluating the effectiveness of the VI re-ID algorithm compared to other widely used measure. Calculation of the average precision

$$AP = \sum_{i=1}^{M_G} Q(i)\Delta R(i) \tag{3}$$

Where $M_G$ is the number of retrieved images, $Q(i)$ represents the precision at the $i-th$ recall level, and $\Delta R(i)$ represents a binary indicator indicating whether the $i-th$ retrieved sample is relevant. The following equation determines the overall mean average precision score:

$$mAP = \frac{\sum_{j=1}^{M_Q} APC_j}{M_Q} \tag{4}$$

Where $M_Q$ is the total number of query samples, and $APC_j$ is the Average Precision for the $q-th$ query sample.

*3) Mean Inverse Negative Penalty(mINP):* When the target appears in the gallery set at multiple time periods, the investigation effort of the examiners is impacted by the most challenging accurate match. To solve this problem, Mang et al. [3] devised an efficient metric called a negative penalty (NP), which evaluates the cost of locating the most difficult right match as

$$NP_i = \frac{R_i^{hard} - |G_i|}{R_i^{hard}} \tag{5}$$

where $R_i^{hard}$ reflects the rating of the toughest match, and $|G_i|$ indicates total accurate matches for query i. As expected, a lower NP indicates superior performance. Specifically, to utilize INP, the inverse operation of NP is consistent with CMC and mAP. When taken as a whole, the average INP of all queries looks like

$$mINP = \frac{1}{n}\sum_i (1 - NP_i) = \frac{1}{n}\sum_i \frac{R_i^{hard}}{|G_i|} \tag{6}$$

Efficiency and smooth incorporation into the CMC/mAP calculation process characterize the process of the mINP computation. For mAP or CMC ranking, mINP ensures that simple matches do not take precedence. For example, the gap between the mINP values for large and small galleries would be smaller for the former.

## IV. Loss Functions of Person Re-ID

A huge proportion of person samples within the person re-ID databases have highly similar appearances, which are brought about by subjective and contextual variables. The variations between the vectors of feature values of the whole image may be little, and in most cases, only local, minute characteristics may be used to discriminate between them. For this reason, various weights must be given to these features. The selection of a loss function plays a crucial role in the training process of transformer-based models for person re-ID [62]. The main objective of the loss function in person re-ID is to acquire an appropriate embedding space whereby the picture representations exhibit distinct separation for persons belonging to different identities, while being closely clustered for individuals of the same identity. The use of transformer-based models is suitable for this particular purpose due to their ability to effectively capture intricate linkages and dependencies within the dataset. Some of the loss functions are explain bellow.

### A. Identity Loss

To classify images with diverse identities and boundaries, for instance, Hao et al. [63] used a Sphere Softmax procedure to train an embedding of a hypersphere, limiting both the within-modality and between-modality variations on this manifold. The multiple-category cross-entropy loss in the person re-ID task is expressed as

$$\mathcal{L}_{id}=-\sum_{i=1}^{K} q\left(x_i\right) \log p\left(y_i \mid x_i\right) \tag{7}$$

Where $x_i$ is the given input image with label $y_i$, $p(y_i \mid x_i)$ is predicted probability, $K$ is the ID category for the training samples in a batch and the label of the sample image of $x_i$ is denoted by $q(x_i)$.

### B. Contrastive Loss

More intelligent use of contrastive loss is beneficial for cross-modality re-ID. Ye et al. [64] proposed a cross-modality hierarchical learning approach by mapping two disparate modalities into a common modality.

$$\mathcal{L}_c=y.d\left(x_i-x_j\right)^2+(1-y)\left[m-d\left(x_i-x_j\right)^2\right]_{+} \tag{8}$$

Where $x_i$ and $x_j$ are two input images of twin networks, while $d(x_i-x_j)$ indicates the correlation between the two images, $y$ is the binary label indicator ($y=1$ if $x_i$ and $x_j$ belong to same identity and $y=0$, otherwise) and $m$ is margin parameter.

### C. Triplet Loss

To account for the dissimilarities between the two sets of information, they devised a bidirectional dual-constrained top-ranking technique that imposes both inter- and intra-modality ranking restrictions. By combining triplet loss with soft margin loss, Hu et al. [65] created a maximum intraclass triplet loss for a cross-modality re-ID. The triplet loss can be expressed as

$$\mathcal{L}_{\text{trip}}=\left[m+d\left(y_a, y_p\right)-d\left(y_a, y_n\right)\right]_{+} \tag{9}$$

where $m$ is margin parameter, $y_a$, $y_p$, and $y_n$ represent anchor, positive, and negative images, respectively and $d(.)$ measures the euclidean distance between two samples.

### D. Quadruplet Loss

The triplet loss may be improved by adding a negative sample picture $y_{n2}$, creating a quadruplet loss if the negative samples $y_{n1}$ and $y_{n2}$ both contain unique pedestrian IDs [66]. The quadruplet-loss function is expressed as follows:

$$\begin{aligned}\mathcal{L}_{\text{quad}} =& \left[m_1+d\left(y_a, y_p\right)-d\left(y_a, y_{n1}\right)\right]_{+}+ \\ & \left[m_2+d\left(y_a, y_p\right)-d\left(y_{n1}, y_{n2}\right)\right]_{+}\end{aligned} \tag{10}$$

Here, $m_1$ and $m_2$ are unique thresholds for training. There is no difference in the anchored image $y_a$ between the positive and negative sample pairs.

### E. Center Loss

To learn modality-shared information, Zhu et al. [67] first created a loss of heterocenters by restricting the center of the classroom distances across two heterogeneous modalities.

$$\mathcal{L}_{\text{center}}=\frac{\lambda}{2} \sum_{i=1}^{N}\left\|x_i-c_{y_i}\right\|_2^2 \tag{11}$$

where $N$ is the number of training samples, $c_y$ is the learned center for class $y_i$ and λ is the momentum term.

### F. Softmax Loss

The softmax loss is a popular choice when combining a classification task with person re-ID. The model generates a high probability of correctly identifying a person owing to softmax loss [68].

$$\mathcal{L}_{\text{softmax}}=-\frac{1}{M} \sum_{i=1}^{M} \log \frac{e^{c_{b_i}}}{\sum_{j=1}^{C} e^{c_j}} \tag{12}$$

where $M$ is the number of training samples and C is the number of classes. $c_j$ is the activation of the $j-th$ neuron in a fully connected layer with weight vector and bias.

### G. Modality-Aware Enhancement Loss

The modality-aware enhancement loss (MAE) loss [69] acts on retrieved features after performing the "Batch Normalization (BN)" that is mainly employed for testing process. Traditionally, the MAE loss comprises of two loss as centre and ID loss, which is used to get intra-class and provide inter-class features. The MAE loss is computed as follows.

$$Lo^{MAE}=Lo^{Cen}+Lo^{ID} \tag{13}$$

The modality-aware centre and ID loss is signified as $Lo^{Cen}$ and $Lo^{ID}$. The selection of the loss function is influenced by the requirements of the person performing the re-ID task, availability of training data, and model architecture. Researchers have frequently experimented with a variety of loss functions to determine which one provides the greatest performance for a given scenario.

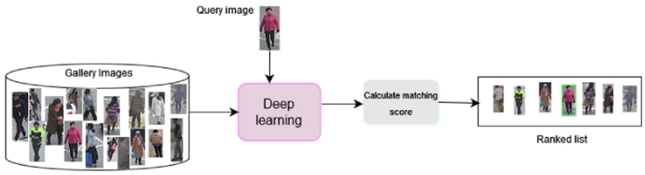

Fig. 4. Person re-identification system based on deep learning techniques.

## V. General Architecture of Person re-ID

The objective of person Re-ID is to locate a person image (probe image) in every gallery image, where the gallery contains a collection of photographs that were taken by various cameras. In this paper, we divided the whole person re-ID researches in two parts one is traditional person re-ID and another part is transformer- based person re-ID. In traditional person re-ID section, we addressed convolutional neural network and traditional deep learning based person re-ID researches. In transformer based we addressed person re-ID researches where transformer based methods used for person re-ID.

### A. Traditional Person re-ID

Recently, deep learning techniques and distance measurements have been combined in a single framework to determine whether or not two photos belong to the same individual (Fig. 4). To do this, either new deep neural networks are created or current deep learning architectures are modified. To handle posture variations for person re-ID, [70] suggested an effective hierarchical indexing and retrieval architecture and global local alignment descriptor. In [71] a comprehensive framework was introduced and trained end-to-end. The framework aims to effectively rank the affinities between the probe and gallery as well as gallery to gallery. In [72] authors addressed the issue of pose variations using a unique unsupervised re-ranking framework, which enabled the learning of both fine and coarse information. The variations in pose difficulties and oversights in pose estimation were effectively addressed by the pose invariant and embedding(PIE) framework [73]. This was achieved using a Pose Box Fusion (PBF) and modulePoseBox alignment module. In [74], a pose-aware multishot matching technique was introduced to cluster the provided photos. In their study [75], an end-to-end attention method was introduced, which incorporated an activation penalty to facilitate the learning of less active and diverse areas. Similarly, [76] introduced a pose-guided framework that incorporated attention mechanisms to address issues related to misalignment and noise elimination. The preservation of global structural information was achieved using an attention mechanism for the purpose of learning contextual information, as discussed in [77]. The authors [78] proposed a comprehensive framework that integrates many components including a visibility prediction model, part visibility, and pose-guided attention models. This framework utilizes a graph-matching approach to achieve its objectives. The identification of particular view angles or the gathering of view-generic or view-invariant person characteristics may significantly contribute to a reduction in intraclass distance and an increase in intraclass distance. Chen et al. [79] suggested an asymmetric distance learning model as a solution for dealing with the challenging issue of multi-camera view. Wu et al. [80] suggested a metric learning approach that is invariant to perspective. This was achieved by incorporating alignment information in advance from the training data and using handcrafted features to represent persons. The authors of [81] introduced a technique for multimetric learning. Cross-view quadratic discriminant analysis was used to determine the relative values of each feature. In [82] authors introduced an unsupervised method to obtain asymmetric cross-view human images. To obtain knowledge on the characteristics that remain consistent regardless of the viewpoint in the study conducted by Chen et al. [83], a joint training framework was proposed by integrating contrastive learning and a generative adversarial network (GAN). The generation of unique views was achieved using a mesh-based view generator.

In instances where a person's face is occluded, the inclusion of extraneous information within the features taken from the whole picture may result in erroneous outcomes if the model lacks the ability to distinguish between the occluded area and the region occupied by the person. Considering a part-based CNN model that uses part-to-part matching, particular attempts at occlusion [84], [85] have improved the performance. The use of attention processes in contemporary person re-ID models was intended to improve emphasis on non-occluded areas. Ma et al. [86] addressed the issue of occluded areas by proposing the use of transformer based on pose-guided relational, which effectively generated a part-aware solution. In [87] authors used an encoding technique that transformed images into an ordered set, thereby creating an extensive vector capable of effectively addressing occlusions.

The optimization of attention-based learning for both spatial and temporal information was achieved by the three highest-performing algorithms [88], [89] and [90] to successfully deal with occlusions. The mechanism of attention was used in a hierarchical fashion in the research conducted in [88] to gather temporal information in both the long and short terms. To learn the spatiotemporal representation via different learning channels, a factorization attention unit was created in [89]. Another prevalent method [90] added temporal and spatial memory networks to enhance individual-level properties. A re-ID solution based on Siamese network was given by the authors [91] to do matching of cross-camera and extract human features at various scales. The scalability issue was addressed by a coarse-to-fine model [92] that learned sub-maps of 3D discriminative features of various sizes. Zhou et al. [93] provided a novel approach for feature extraction at many scales, referred to as omni-scale feature extraction. This method incorporates diverse receptive fields and replaces ordinary convolution with point- and depth-wise convolutions. In the study conducted by Yan et al. [94], multiscale cross attention was introduced. The purpose of this model is to acquire discriminative information pertaining to the various body parts of a specific individual from several perspectives. by incorporating a multi-scale layer. Qian et al. [95] addressed the issue of multi-scale complexity in their model. This layer enables multi-scale feature learning at both the global and local levels. Additionally, they introduced an attention learning layer based on the leader that learns the most effective weight assigned. An alternative approach to a pyramid of multi-scale attention was introduced in [96]. This method uses spatial attention and channel-wise modules that enable learning of attention characteristics at several local levels. In [97], a dual-refinement methodology was introduced to mitigate the effects of noisy labels. This strategy involved refining pseudo-labels via offline clustering and performing feature refinement during the online training phase. To address the issue of pseudo-label noise, Zhang et al. [98] proposed a method that involves refining pseudo-labels using clustering consensus. In [99], a "divide-and-conquer" approach was employed to address the domain gap in factor-wise style transfer, building on the work of Liu et al. [100] on adaptive style transfer and Zhou et al. [101] on CycleGAN. Similarly, Tang et al. [102] employed the CycleGAN framework to facilitate the transfer of label information from a labeled domain to an unlabeled domain. In addition, they utilized the maximum mean discrepancy as a foundational measure to effectively minimize the dissimilarity between distributions. A different technique based on CycleGAN, as proposed by Deng et al. [103], introduced the concept of "learning via translation" and extended its application to various domains in an unsupervised fashion. The study conducted by Li et al. [104] utilized a novel attention network to simultaneously acquire attention-based pixel representations in both hard and soft regions. A different formulation for attention aggregation was proposed in [105] to address the dynamic representation of an identity in a query image. The issue of varying the camera viewpoints in videos collected from different cameras was addressed in [106].

All of the extant research are founded upon deep learning-based methodologies and take into account one or two specific problems. Nevertheless, these traits lack discriminative power and do not remain consistent in the presence of both inter-class and intra-class changes. Hence, this assessment focuses on the analysis of transformer-based approaches, since they have shown substantial advancements in person re-ID systems, equivalent to traditional deep learning techniques. Transformer based methods possess the capability to collect global context through their self-attention mechanisms. This is particularly advantageous in the context of person re-ID, since the ability to identify overarching patterns such as overall attire, posture, or carrying goods is just as essential as identifying specific local information, such as faces or brands. In the person re-ID context, it is observed that some characteristics possess more discriminative power than others. Using a dynamic self-attention mechanism, the model can selectively emphasize certain discriminative traits, thereby enhancing its overall performance. The use of transformers facilitates the establishment of interactions among all possible combinations of features, thereby allowing the model to effectively represent the intricate connections present within the data. This may be advantageous in situations in which a combination of numerous features delineates an individual's identity. The use of multihead attention and the stacking of numerous transformer layers facilitate the acquisition of hierarchical representation learning. This enables the model to acquire knowledge related to low-level characteristics, such as edges or textures, and high-level concepts, such as posture.

### B. *Person re-ID Based on Transformer Models and Its Variants*

The transformer model used in computer vision was initially studied by Han et al. [107] and Salman et al. [108]. Lu et al. [109] introduced a swin transformer as the network backbone for the mask transformer model, which detects and classifies images collected by unmanned aerial vehicles. In their research, Lee et al. [110] introduced a novel approach called the attribute debiased vision transformer (AD-ViT), which aims to provide explicit guidance for the acquisition of identity-specific characteristics. Yulin et al. [111] introduced a transformer encoder-decoder architecture based on part-aware for occluded person re-ID via various component discoveries. Xu [112] suggested a transformer-based feature extractor to learn distinguishing characteristics using the self-attention method to address occlusion in person re-ID. To clearly untangle semantic components (such as the human body or joint parts) and selectively match non-occluded portions, Wang et al. [113] developed a posture-guided transformer-based feature-disentangling technique. A disentangled representational learning network was reported by Jia et al. [16] which might handle occluded re-ID without the need for further supervision or exact person picture alignment. The authors [114] developed a framework for re-identifying occluded people based on partial feature transformers, and used patch full-dimension augmentation, fusion, reconstruction, and slicing of spatial modules to boost vision transformer effectiveness. A revolutionary cloth-changing-aware transformer was presented in [115] to learn identity-relevant stimuli. The transformer encoder employs Cloth Information Embedding (CIE) to encrypt clothing information. It also automatically concentrates on the identity-discriminative feature and imposes intraclass constraint learning to draw the global tokens from the encoder closer.

A vision transformer-based approach was suggested by Bansal et al. [116] to deal with cloth-changing-based long-term person re-ID. In [117] and [118], the authors provided approaches for incorporating transformer networks in person re-IDs on a video. For video-based re-ID, Zhang et al. [119] suggested a transformer that allows for the separation of time and space. With a constraint on cross-modal consistency, [20] proposed a dual-attention collaborative learning technique that integrates spatially attentive deep features to collect additional data for multiple classifiers. In [120], a robust re-ID baseline was built on top of a pyramidal transformer with a conv-patchify operation, termed PTCR, which inherits the advantages of both the CNN and Transformer. The pyramidal structure captures multi-scale fine-grained features, whereas conv-patchify enhances the robustness against translation. Vision transformers can simulate long-range dependence by utilizing self-attention, making them useful in computer vision problems. Li et al. [121] proposed the ViTAE transformer, which uses a reduction cell for multi-scale features and a normal cell for localization, using the two IBs.

Transformers make re-ID successful in many applications. Existing works employ the transformer's highest-level information to distinguish a few pieces or locations. A few focused pieces in the re-ID cannot differentiate an inquiry individual under multiple scenarios and camera angles. Tan et al. [122] proposed a high-performing Multi-level Feature aggregate transformer for person re-identification (MFAT). To learn semantically aware transferable modality properties, Chen et al. [19] introduced a transformer network. In [123] the authors suggested a Spectrum-Insensitive Data Augmentation (SIDA) method, which successfully reduces disruption in the visible and infrared spectra and compels the network to learn properties unrelated to the spectrum.

Video-based person re-ID, a primary IoT application, identifies the same person in multiple video sequences using non-overlapping cameras. Wang et al. [88] proposed Pyramid Spatial-Temporal Aggregation (PSTA) framework aggregates frame-level features and fuses hierarchical temporal features into a final video-level representation, allowing for both short-term and long-term temporal information to be exploited. Tang et al. [102] presented a unique Multi-Stage Spatial-Temporal Aggregation Transformer (MSTAT) with two proxy embedding modules to extract local characteristics and global identity information for person re-ID. In [124], the authors proposed a novel Spatiotemporal Interaction Transformer Network (SITN) to solve the problem of learning frame-level spatial representations from temporal cues. Liu et al. [125] proposed a deeply coupled convolution-transformer (DCCT), a novel spatial-temporal complementary learning framework for high-performance video-based person re-ID.

Visible-infrared person re-ID is a tough re-ID job that retrieves and matches photographs of the same identity across diverse modalities. Liang et al. [126] proposed a new cross-modality transformer-based method (CMTR) to solve the visible-infrared person re-ID task by explicitly mining the information for each modality and generating better discriminative features. In [127], the authors designed a TransVI network with a two-stream structure to capture modality-specific representations and learn multimodality sharable knowledge to bridge the modality gap and assess the overall distance distribution discrepancy. Structure-aware positional transformer (SPOT) networks were developed by Chen et al. [19] to learn semantic-aware sharable-modality characteristics using structural and positional information. In this review, we have summarized some of the recent transformer-based methods for person re-ID on different single-modality and cross-modality datasets in Table V.

Researchers have explored different transformer variants in the context of person re-ID to adapt the transformer architecture to this specific task. Some notable transformer variants used in person re-ID are as follows:

*1) Vision Transformer Based Person re-ID:* The original vision transformer is excellent at gathering long-range associations between patches while ignoring local feature extraction, projecting the 2D patch to a vector with a fundamental linear layer. Researchers have recently focused on increasing the modeling capability of local information [128], [129], [130]. TNT [128] employs a new transformer-in-transformer design to represent the interaction between sub-patches and communicate patch-level information. Cross-window connections are made possible by a Shuffle Transformer [131], [132], which also uses a spatial shuffle operation rather than shifting window partitioning. When regional tokens are created from an image using RegionViT, local tokens receive attention from the global community [130], which also creates regional and local tokens. In addition to local attention, several additional efforts, such as T2T [133], have suggested enhancing local information by aggregating local features. These pieces highlight the value of both local and international information communications in vision transformers. One crucial future path is the re-identification of a person by utilizing these local and global features. To perform image classification, the Vision Transformer [134] is a standard transformer that may be used on picture sequences instead of intermediate steps. It closely adheres to the original design of the transformers.

The original image $\mathcal{X} \in \mathbb{R}^{w\times h\times c}$, with dimensions of $(h, w)$, is transformed into a collection of 2D patches, $\mathcal{X}p \in \mathbb{R}^{n\times(p^2.c)}$, where c is the number of channels and $(h, w)$ is the image's resolution. Similarly, the pixel dimensions of the image patch are denoted by $(p, p)$. In this case, the transformer's efficient sequence length is denoted by $n = \frac{hw}{p^2}$. The transformer's output is termed patch embedding [135], and it is a trainable linear projection that maps each vectorized route to the model dimension d. For pretraining, the ViT model requires a large dataset. To address this deficiency, Touvron et al. [136] proposed the DeiT framework, which adds a teacher-student technique, particularly to transformers, to accelerate the training of ViT without first collecting massive amounts of data. Arnab et al. [137] provided several effective variations in their model that factorize the spatial and temporal dimensions of the input to manage the lengthy sequences of tokens observed in the video.

Recent studies have explored the adaptation of transformer models, initially designed for natural language processing, to address the challenges of person re-ID. In [107] and [108], the authors assessed the transformer applications in computer vision. Researchers have recently used it for several visual tasks, such as object recognition [138], image classification [134], and semantic segmentation [139], and found it to be more effective than convolutional neural networks.

Vision transformers replace CNNs for object recognition and person re-identification. Vision transformers generate global classification and local tokens using the picture area information. Sharma et al. [140] introduced the locally aware transformer-based person re-ID. ViT has shown promising results for person re-ID, demonstrating the effectiveness of transformer-based models for this task.

*2) Swin Transformer Based re-ID:* The Swin transformer, introduced in 2021, is a variant designed to handle large-scale images. A hierarchical transformer with shifted-window representation addresses these issues [129]. In [141], the authors offered a two-fold loss swin transformer to focus on a pedestrian's semantic information and retain helpful background information, thereby decreasing background appearance interference. In [142] authors proposed an innovative and robust person re-ID model that leverages the strengths of both CNN and transformers. Swin Transformer architecture, which allowed for effective integration of these two techniques. In order to augment the capacity for capturing long-range dependencies, they proposed the introduction of the Adaptive Split Self-Attention (ASA) module.

TABLE V
Summery of Transformer-Based Person Re-Id

| Reference | Approach | Contribution | Dataset | Published year |
|---|---|---|---|---|
| He et al. [13] | TransReID | To learn a stable feature representation for a single-modality re-ID, this study presents a pure transformer-based architecture. | DukeMTMC-reID, Market-1501, MSMT17, Occluded-Duke, VeRi-776, and VehicleID | 2021 |
| Zhang et al. [118] | STT | Developed a spatial-temporal transformer for video-based re-ID information extraction. | DukeMTMC-VideoReID, MARS and LS-VID | 2021 |
| Shenqi et al. [146] | TMM | TPM and PMG modules are used. TPM first adaptively allocates semantic object patch tokens to identical parts. PMG then creates numerous non-overlapping masks for robust part division from these similar parts. | Market-1501, CUHK03, DukeMTMC-ReID and MSMT17 | 2021 |
| Zhu et al. [14] | AAformer | It uses the ViT backbone and ”part token” vectors to acquire part representations and align self-attention. | Market-1501. DukeMTMC, CUHK03-NP, MSMT17 | 2021 |
| Hao et al. [147] | SSPT | Initially, we examined self-supervised learning approaches utilizing a vision transformer pre-trained on unlabeled human photos. They showed that it performs better on re-ID tasks than imageNet-supervised pretraining models. | Market-1501, MSMT17 | 2021 |
| Guowen et al. [143] | HAT | Intensely Supervised Aggregation to repeatedly aggregate hierarchical CNN backbone characteristics. Introduced Transformer-based feature calibration to use low-level detail information as the global prior for high-level semantic knowledge. | Market-1501. DukeMTMC, CUHK03-NP, MSMT17 | 2021 |
| Chen et al. [148] | ResT | Using the local perception and shared parameters of CNNs and the Transformer block model’s long-range dependent features, a robust hybrid re-ID architecture was developed for feature extraction with discriminatory potential. | Market-1501, DukeMTMC-ReID, Occluded DukeMTMC-ReID | 2022 |
| Haoyan et al. [15] | Denseformer | Presented a dense transformer system for the person re-ID using class tokens to link each layer. | Market-1501, DukeMTMC, MSMT17, and Occluded-Duke | 2022 |
| Wang et al. [113] | PFD | Initially, ViT extracts patch characteristics. Pose-guided Feature Aggregation (PFA) utilizes the matching and distributing process to separate posture and patch information. Finally, the transformer decoder indirectly enhances disentangled body component characteristics with learnable semantic perspectives. | Occluded Duke, Occluded REID, Market-1501, DukeMTMC-reID, MSMT17 | 2022 |
| Jiachen et al. [145] | DNA | The global tokens of two branches are averaged to provide a global feature. In contrast, the local tokens of each component are reshaped and evenly partitioned into several stripes to generate part-level features. | Market1501, DukeMTMC-reID, and MSMT17 | 2023 |
| Wang et al. [149] | TCiP | Combines clothes-changing-resistant local properties. Patch features are extracted using a transformer network for outstanding performance in numerous visual applications. Patch picker is an innovative patches sequence recombine module based on human parsing. | Market-1501,MSMT17. | 2023 |
| Chen et al. [19] | SPOT | It relies on attended structural representation(ASR) and transformer-based part interaction. ASR dynamically selects discriminative appearance regions utilizing structure information from each modality’s modality-invariant structure feature. | SYSU-MM01 | 2022 |
| Wang et al. [150] | AlignGAN | This approach utilizes both pixel and feature alignment simultaneously. The model under consideration comprises a pixel, feature, and joint discriminators. | SYSU-MM01, RegDB | 2019 |
| Feng et al. [21] | CMIT | This study presents a matching approach for occlusion conditions utilizing LFEM and modality information fusion modules (MIFM). Transformer learns modality features and alters patch priority to match non-occluded area qualities in LFEM. | SYSU-MM01, RegDB | 2022 |
| Li et al. [20] | DAC | Channel and spatial attentive deep features enhance multiple classifiers with cross-modal consistency in dual-attention collaborative learning. Consistency is needed for multiple-classifier cross-modal identification. | SYSU-MM01 | 2022 |
| Jiang et al. [22] | CMT | Transformer encoder-decoder architecture substitutes modality-specific information in the modality-level alignment module. The query-adaptive feature modulation instance-level alignment module modifies sample features. | SYSU-MM01, RegDB | 2022 |
| Liu et al. [123] | MCP | Two Transformer-based feature extraction networks separately extract global and non-occluded person region features; A multi-headed self-attention method learns the distribution of common non-occluded human areas, and then the Minimized Character-box Proposal (MCP) is used to create correct shared crops. | Occluded-Duke, Occluded-REID, Partial-REID, Partial-iLIDS. | 2023 |

The ASA module enhances the global modeling of inter-regional relationships by using self-attention on the adaptable partitions of the input. The Swin transformer has demonstrated strong performance in image recognition tasks and has the potential to be adapted for person re-ID.

*3) Hierarchical Transformers Based re-ID:* Hierarchical Transformers address the hierarchical structure of person images. They decomposed an image into different levels, such as body, part, and region levels, and applied transformer models at each level. This allows the model to capture both local and global contextual information. Zhang et al. [143] present a high-performance learning system called Hierarchical Aggregation Transformer (HAT) for image-based person re-ID using CNNs and Transformers. Hierarchical Transformers have improved the handling of the challenges of person re-ID, particularly in capturing fine-grained details and distinguishing subtle differences.

*4) Multi-Granularity Transformer Based re-ID:* Multi-Granularity Transformers incorporate multiple scales or feature granularities in person re-ID. By considering multiple granularities, the models can effectively handle variations in body scale, partial occlusions, and other challenges in person re-ID. In [144]

TABLE VI
COMPARISON OF RESULTS FOR SOME POPULAR DATASETS BASED ON PERSON RE-ID METHODS

| Method | Backbone | Market-1501 | | DukeMTMC | | Occluded-Duke | | MSM17 | | CUHK03 | | Reference |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Rank-1 | mAP | Rank-1 | mAP | Rank-1 | mAP | Rank-1 | mAP | Rank-1 | mAP | |
| RGA [77] | CNN | 96.1 | 88.4 | | | | | | | 81.1 | 77.4 | CVPR,20 |
| SAN [153] | CNN | 96.1 | 88.0 | 87.9 | 75.7 | | | 79.2 | 55.7 | | | AAAI,20 |
| SCSN [154] | CNN | 95.7 | 88.5 | 91.0 | 79.0 | | | 83.8 | 58.5 | | | CVPR,20 |
| ISP [155] | CNN | 95.3 | 88.6 | 89.6 | 80.0 | 62.8 | 52.3 | | | | | ECCV,20 |
| PAT [111] | CNN | 95.4 | 88.0 | 88.8 | 78.2 | 64.5 | 53.6 | | | | | CVPR,21 |
| HLGAT [156] | CNN | 97.5 | 93.4 | 92.7 | 87.3 | | | | | 83.5 | 80.6 | CVPR,21 |
| TransReID [13] | DeiT-B/16 | 95.0 | 88.4 | 91.1 | 81.9 | 66.4 | 58.1 | 84.53 | 66.2 | | | ICCV,21 |
| TransReID [13] | ViT-B/16 | 95.2 | 88.9 | 90.7 | 82.0 | 66.4 | 59.2 | 85.3 | 67.4 | | | ICCV,21 |
| MGN [146] | Resnet-101 | 96.0 | 90.3 | 91.3 | 82.1 | | | 82.9 | 62.7 | 79.2 | 75.7 | ICCV,21 |
| SSPT [147] | ViTI -B | 96.7 | 93.2 | | | | | 89.5 | 75.0 | | | arXiv-21 |
| AAformer [14] | ViT-B/16 | 95.4 | 87.7 | 90.1 | 80.0 | | | 83.1 | 62.6 | 77.6 | 74.8 | arXiv,21 |
| Denseformer [15] | Transformer | 95.0 | 88.1 | 90.0 | 81.3 | 63.8 | 55.6 | 84.1 | 65.3 | | | IETCV,22 |
| ResT-ReID [148] | Transformer | 95.3 | 88.2 | 90.0 | 80.6 | 59.6 | 51.9 | | | | | PR,22 |
| DRL-Net [16] | Transformer | 94.7 | 86.9 | 88.1 | 76.6 | 65.0 | 50.8 | | | | | TMM,22 |
| TMGF | Transformer [146] | 96.3 | 91.9 | 92.3 | 83.1 | | | 88.2 | 70.3 | | | WACV,23 |
| X-ReID [17] | Transformer | 94.9 | 65.1 | | | | | 84.0 | 65.1 | | | ArXiv,23 |
| DAAT [18] | Transformer | 95.1 | 88.8 | 90.6 | 82.0 | 63.3 | 57.1 | | | | | IVC,23 |

presented a novel approach that utilizes a multi-granularity temporal convolution network and mutual distance-matching measurement. The objective was to mitigate the effects of both intra-sequence and inter-sequence variation. In the feature-learning process, we use a hierarchical approach to describe various temporal granularities. This was achieved by layering temporal convolution blocks with varying dilation factors. While in [145], the authors extracted multi-grained features from a pure transformer network to solve the label-free but complicated unsupervised re-ID problem. We use a modified ViT to establish a dual-branch network.

These transformer variants demonstrate the adaptability of the transformer architecture for person re-ID. By leveraging self-attention mechanisms and capturing long-range dependencies, these models have demonstrated the potential to improve feature learning, handle global contexts, and enhance discriminative representations for person re-ID. Further research and advancements in transformer-based models are expected to continue to push the boundaries of the person re-ID performance.

## VI. COMPARISON AND DISCUSSION

Owing to the development of transformer-based methods, person re-ID has become significantly more prominent in computer vision and the rising need for intelligent video monitoring.

### *A. Person re-ID Based on Single Modality Datasets*

This section compares the modern techniques applied to various datasets, such as Market-1501, MSM17, DukeMTMC, Occluded-Duke, CUHK03, SYSU-MM01, and RegDB. Table VI presents the experimental outcomes for person re-ID based on conventional DL methods and transformer based techniques using Market-1501, DukeMTMC, Occluded-Duke, MSM17, and CUHK-03. We identified two significant difficulties inadequately addressed within the re-ID by examining CNN-based approaches. First, it is essential for re-IDs to use several structural patterns on a global scale [77]. Considering the Gaussian distribution of practical receptive fields [151], CNN-based algorithms concentrate primarily on small discriminative regions. Second, fine-grained characteristics and detailed information are essential. Nevertheless, the down-sampling operators (such as pooling and strided convolution) of the CNN diminish the granularity of the generated feature maps in terms of space, which has a significant impact on the discriminating capacity [84], [152].

### *B. Person re-ID Based Cross-Modal Datsets*

Transformer-based approaches have shown significant progress in NLP and CV. Numerous academics have conducted person re-ID using transformer-based networks to concentrated large discriminative regions. Table VII and Table VIII also shows the experimental findings of several conventional DL methods and recent transformer-based methods on the SYSU-MM01 and RegDB visible-infrared datasets respectively.

However, most existing models largely focus on person re-ID using visual images exclusively and within a single modality. There is a requirement for a person search that encompasses both daytime and nighttime activity. However, the limited efficacy of these techniques under low-light conditions imposes constraints on their use in continuous surveillance systems operating around the clock. Compared to visible cameras, infrared (IR) cameras are capable of acquiring sufficient data from scenes under low-light conditions. Additionally, security systems commonly incorporate dual-mode cameras to enable automatic transitions in response to variations in illumination conditions. To mitigate huge modality gap between different modalities transformer based methods are very effective on person re-ID. In these context, our review will be helpful for future person re-ID researchers.

### *C. Benefits and Limitations of Various Methods*

For the several deep learning methods involved in the previous article, the advantages and disadvantages of these methods are summarized below:

TABLE VII
COMPARISON OF RESULTS FOR SYSU-MM01 DATASET ON TRANSFORMER-BASED PERSON RE-ID

| Methods | All search | | Indoor search | | Reference |
|---|---|---|---|---|---|
| | Rank-1 | mAP | Rank-1 | mAP | |
| Deep-learning methods | | | | | |
| JSIA [158] | 38.10 | 36.90 | 43.80 | 52.90 | AAAI-2020 |
| Xmodal [159] | 49.92 | 50.73 | | | AAAI-2020 |
| CMSP [160] | 43.56 | 44.98 | 48.62 | 57.50 | IJCV-2020 |
| DDAG [161] | 54.75 | 53.02 | 61.02 | 67.98 | ECCV-2020 |
| cm-SSFT [162] | 61.60 | 63.20 | 70.50 | 72.60 | CVPR-2020 |
| AGW [3] | 47.50 | 47.62 | 54.17 | 62.97 | TPAMI-2021 |
| NFS [163] | 56.91 | 55.45 | 62.79 | 69.79 | CVPR-2021 |
| CICL [164] | 57.20 | 59.30 | 66.60 | 74.70 | AAAI-2021 |
| DML [165] | 58.40 | 56.10 | 62.40 | 69.60 | TSCVT-2022 |
| HAT [166] | 55.29 | 53.89 | 62.10 | 69.37 | TIFS,2021 |
| DLS [167] | 48.80 | 49.00 | | | TMM,2021 |
| WIT [168] | 59.20 | 57.30 | 60.70 | 67.10 | NC,2021 |
| DFLN-ViT [23] | 59.84 | 57.70 | 62.13 | 69.03 | TMM,2022 |
| CMDSF [169] | 59.97 | 57.28 | | | KBS,2022 |
| DTRM [170] | 63.03 | 58.63 | 66.35 | 71.76 | TIFS,2022 |
| TSME [171] | 64.23 | 61.21 | 64.80 | 71.53 | TCSVT,2022 |
| SFANet [172] | 65.74 | 60.83 | 71.60 | 80.05 | TNNL,2021 |
| DART [173] | 68.72 | 66.29 | 72.52 | 78.17 | CVPR,2022 |
| M2FINet [174] | 74.73 | 68.96 | 76 39 | 79.20 | CVIU,2023 |
| Transformer based methods | | | | | |
| SPOT [19] | 65.34 | 62.25 | 69.42 | 74.63 | TIP,2022 |
| DAC [20] | 57.33 | 54.49 | 62.74 | 68.49 | SPIC,2022 |
| CMIT [21] | 70.94 | 65.54 | 73.28 | 77.18 | TMM-2022 |
| CMTR [22] | 62.58 | 61.33 | 67.02 | 73.78 | arXiv-2021 |
| DFLN [23] | 59.84 | 57.70 | 62.13 | 69.03 | TMM-2022 |

*1) Manually Built Features Model:* Benefits: The structure of this model is simple and easy to understand and label information required to this model. Limitations: The scalability of hand-crafted features may be limited when dealing with large or high dimensional datasets due to the impractically of producing and managing an extensive amount of features. Domain expertise and physical effort are needed to create successful hand-crafted features. This requires time and may not enhance model performance.

*2) Metric Learning Model:* Benefits: Improved performance in tasks that rely on similarity. The ability to withstand fluctuations and disruptions in data. The mitigation of semantic gaps within feature representations. Limitations: Training requires labeled data, which may be expensive and scarce. Metric learning models might suffer from noisy or poor data.

*3) Part-Based Deep Model:* Benefits: Part-based models divide things into pieces for more precise identification. This helps with complicated forms and deformations. Part-based models can record object posture and perspective alterations, making them more flexible. Limitations: Setting up, training, and tweaking part-based models may be difficult. All item categories may not have enough labeled data for these models to train.

*4) GAN-Based Model:* Benefits: GANs can create high-quality data samples that match training data. This is useful for data enrichment, rare event training, and image, audio, and text generation. GANs may represent and produce data without labels or supervision. They find data patterns and structures. Limitations: GANs may experience mode collapse, generating a restricted collection of comparable samples that fail to replicate the training data's variety. Strong GPUs and high memory capacity are needed for GAN training, which might be costly and inaccessible.

*5) Unsupervised Model:* Benefits: Unsupervised models may identify abnormal data, making them useful in anomaly detection, defect monitoring, and quality control. Unsupervised learning may reduce the requirement for human data labeling, saving time and resources. This is especially useful when labeled data is scarce or expensive. Limitations: Unsupervised models may be difficult to evaluate since their learnt representations or clusters may not be easily measurable.Unsupervised models may collect irrelevant data patterns, making them less useful for certain applications.

*6) Supervised Model:* Benefits: Supervised models may forecast accurately on many tasks, making them suited for essential jobs. Domain experts may label training data to include domain-specific knowledge into the model. Limitations: Labeled data for supervised learning is costly and time-consuming, particularly for large or specialized datasets. Label noise may result from annotators' subjective interpretations of data.

The use of fused hand-crafted features in the initial phases of DL is becoming prevalent. Since 2017, there has been a reduction in research focused on approaches associated with hand-crafted features due to the emergence and popularity of alternative DL methods. Consequently, the use of technology related to hand-crafted features has become obsolete in contemporary contexts. The robustness of pedestrian features in the presence of complicated environmental changes is enhanced by the high feature extraction capabilities of CNNs. Following the gradual emergence of deep learning, further developments in this area prompted the development of representation learning and metric learning models. The combination of these two strategies is often used in the advanced stages of re-ID technology development to attain mutually beneficial outcomes. The use of the part-based technique has become more prevalent in the image-based supervised approach. This particular methodology often involves the extraction of local information through posture estimation, feature map division, and attention processes. These local information are subsequently merged with the global features of the pedestrian to provide an invariant representation of the pedestrian. The year 2017 saw the first use of the pedestrian re-ID approach using GAN [87], therefore introducing a new avenue for re-ID research. The GAN-based re-ID model not only addresses the issue of data augmentation but also enables the accomplishment of cross-domain re-ID tasks, hence introducing novel perspectives for unsupervised person re-identification tasks. The implementation of re-ID technology faces significant obstacles. As a result, there has been a progressive emergence of unsupervised approaches in the field. These methods include techniques such as clustering, tracklet analysis, and GANs to accomplish unsupervised person re-ID. Despite some improvements in the generalization of the unsupervised model, its performance remains significantly inferior to that of the supervised technique. This might be attributed to the absence of supervision information.

TABLE VIII
COMPARISON OF RESULTS FOR REGDB DATASET ON TRANSFORMER-BASED PERSON RE-ID

| Method | visible-to-infrared(VI) | | infrared-to-visible(IV) | | Reference |
|---|---|---|---|---|---|
| | Rank-1 | mAP | Rank-1 | mAP | |
| Deep-learning based methods | | | | | |
| JSIA [158] | 48.50 | 49.30 | 48.10 | 48.90 | AAAI-2020 |
| Xmodal [159] | 62.21 | 60.18 | | | AAAI-2020 |
| CMSP [160] | 65.07 | 64.50 | | | IJCV-2020 |
| AGW [3] | 70.05 | 66.37 | | | TPAMI-2021 |
| HAT [166] | 71.83 | 67.56 | 70.02 | 66.13 | TIFS-2020 |
| SFANet [172] | 76.31 | 68.00 | 70.15 | 63.77 | TNNL-2020 |
| CICL [164] | 78.80 | 69.40 | 77.90 | 69.40 | AAAI-2021 |
| NFS [163] | 80.54 | 72.10 | 77.95 | 69.79 | CVPR-2021 |
| WIT [168] | 85.00 | 75.90 | | | NC-2021 |
| MPMN [12] | 86.56 | 82.91 | 84.62 | 79.49 | TIP-2021 |
| DART [173] | 83.60 | 60.60 | 81.97 | 73.78 | CVPR-2022 |
| DML [165] | 77.60 | 84.30 | 77.00 | 83.60 | TCSVT-2022 |
| CMDSF [169] | 84.75 | 77.91 | | | KBS-2022 |
| TSME [171] | 87.35 | 76.94 | 86.41 | 75.70 | TCSVT-2022 |
| M2FINet [174] | 92.84 | 85.37 | 90.96 | 83.64 | CVIU-2023 |
| Transformer based methods | | | | | |
| CMTR [22] | 80.62 | 74.42 | 81.06 | 73.75 | arXiv-2021 |
| SPOT [19] | 80.35 | 72.46 | 79.37 | 72.26 | TIP,2022 |
| DAC [20] | 85.33 | 82.10 | 84.07 | 80.56 | SPIC,2022 |
| CMIT [21] | 88.78 | 88.49 | 84.55 | 83.64 | TMM-2022 |
| DFLN [23] | 92.10 | 82.10 | 91.21 | 81.61 | TMM-2022 |

Transformer models were revolutionized for processing sequences of information in the natural language processing area in 2017 by Vaswani et al. [157] and have become the SOTA approach for various tasks. At the core of the transformer model are self-attention mechanisms, which allow the model to weigh the importance of different words or tokens in a sentence when generating or understanding a sequence. He et al. [3] introduced transformer based on person re-ID in 2021. This method used to solve visible image-based person re-ID not applicable for cross-modal based person re-ID. The performance of transformer-based approaches will be improved, and this will be an important area for future study. The majority of the study added one or two vision-related issues into the transformer-based solutions that have surfaced in recent years in order to improve outcomes. Therefore, in the future, researchers may combine methodologies by comprehending the benefits and drawbacks of different transformer versions, optimize advantages and minimize downsides, and enhance the performance of re-ID models by using transformer-based techniques.

## VII. CHALLENGES AND FUTURE RESEARCH DIRECTIONS

Person re-ID faces several challenges, and researchers are actively exploring future directions to address them. Some key challenges and potential directions in the field are as follows:

### A. Handling Large-Scale Datasets

With the increasing magnitude and heterogeneity of person re-ID datasets, there is a need for models that can effectively process extensive-scale datasets. The architectural design of the transformers enables effective parallel processing, thereby facilitating scalability. This strategy has proven advantageous in scenarios involving vast re-identification datasets or circumstances in which the gathering of detailed features requires the use of high-resolution photos. To effectively manage datasets of significant magnitude, researchers have emphasized the use of person re-ID based on transformer methods.

### B. Real-Time Inference

The actual implementation of real-world person re-ID systems depends heavily on their ability to perform real-time inference. The ongoing problem lies in the optimization of transformer-based models to achieve real-time performance. Possible future research areas could include investigating alterations to model design, optimizing attention processes for improved efficiency, employing knowledge distillation strategies, and leveraging hardware acceleration to enhance the speed and efficiency of inference in person re-ID models.

### C. Robustness to Occlusion and Pose Variations

Person re-ID models must be resilient against occlusion, position changes, and other difficult circumstances that are often present in real-world surveillance settings. The layers of the transformers' self-attention systems enable direct interactions between the levels. This makes it easier to extract features at several levels, thereby capturing both high-level semantics and low-level details. Future research may concentrate on creating transformer-based techniques that successfully manage occlusions, differences in illumination, viewpoints, and other intraclass variances.

### D. Domain Adaptation and Generalization

When used in new domains with distributions that vary from the training data, person re-ID models often perform poorly. Large transformer models that have already been trained on large datasets may be tailored for certain purposes such as person re-ID. Even with little labeled data in the re-ID domain, the rich feature representations gained during pre-training may greatly improve performance. Future research may look at transformer-based domain adaptation methods to increase the capacity of the model to generalize across many domains. These methods narrow the domain gap and enhance the performance of the model when applied to new datasets.

### E. Multi-Modal Person re-ID

Person re-ID shows great promise when different modalities, such as pictures, videos, and text descriptions, are integrated. Transformers may dynamically complement the significance of different input elements owing to their self-attention mechanism. All feature pairs may naturally interact with each other in the transformers. As people may be distinguished by the complicated relationships among traits, this is helpful in person re-ID. Transformers may be made aware of the location of the features using positional encodings. In this situation, researchers have concentrated on investigating cross-modal attention operations, multimodal learning methodologies, and fusion strategies to effectively employ favorable data from various modalities for person re-ID.

These challenges and future prospects demonstrate how person re-ID research is constantly evolving. Researchers may progress this area and create more reliable, effective, and moral person re-ID systems that meet real-world needs by tackling these issues and looking at new approaches. Transformer-based approaches are a great area of study because of their infancy in person re-ID. Future possibilities include more effective transformer designs, hybrid models that combine CNNs and transformers, and unsupervised or semi-supervised learning techniques that use transformers for re-ID.

## VIII. Conclusion

The most modern methods for person re-ID are described in this survey. First, the significance and complexity of this work are discussed. Second, a summary of current re-ID research is provided. We classify these strategies as handcrafted, based on deep learning, or a transformer. We present a comprehensive summary of various widely used re-ID datasets. Finally, we provide a quick summary of how various latest techniques have been performed over the last several years on the re-ID datasets that are available. Transformer-based methods have demonstrated significant potential and promising results in person re-ID tasks. By leveraging their attention mechanisms, sequence modeling capabilities, scalability, and cross-modal fusion, transformer-based models have shown advantages in capturing complex dependencies, handling large-scale datasets, and integrating information from different modalities. Transformer-based methods for person re-ID are still the subject of ongoing research. Future studies can explore different transformer variants, design architectures that explicitly address temporal modeling, incorporate attention mechanisms at different levels, and consider self-supervised learning or unsupervised domain adaptation to improve the performance. We expect that this review will be valuable in future studies.

## Acknowledgment

The authors would like to thank all the participants for their contributions in this work.

## References

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

[2] N. Gheissari, T. B. Sebastian, and R. Hartley, "Person reidentification using spatiotemporal appearance," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2006, pp. 1528–1535.

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

[4] Z. Zhong, L. Zheng, Z. Luo, S. Li, and Y. Yang, "Learning to adapt invariance in memory for person re-identification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 8, pp. 2723–2738, Aug. 2021.

[5] Y. Lin, L. Xie, Y. Wu, C. Yan, and Q. Tian, "Unsupervised person re-identification via softened similarity learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3390–3399.

[6] Z. Zheng, L. Zheng, and Y. Yang, "Pedestrian alignment network for large-scale person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 10, pp. 3037–3045, Oct. 2019.

[7] Y. Jing, C. Si, J. Wang, W. Wang, L. Wang, and T. Tan, "Pose-guided multi-granularity attention network for text-based person search," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 11189–11196.

[8] J. Yang, W.-S. Zheng, Q. Yang, Y.-C. Chen, and Q. Tian, "Spatial-temporal graph convolutional network for video-based person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3289–3299.

[9] G. Chen, J. Lu, M. Yang, and J. Zhou, "Spatial-temporal attention-aware learning for video-based person re-identification," *IEEE Trans. Image Process.*, vol. 28, no. 9, pp. 4192–4205, Sep. 2019.

[10] Y. Liu, W. Zhou, J. Liu, G.-J. Qi, Q. Tian, and H. Li, "An end-to-end foreground-aware network for person re-identification," *IEEE Trans. Image Process.*, vol. 30, pp. 2060–2071, 2021.

[11] H. Zhang, H. Cao, X. Yang, C. Deng, and D. Tao, "Self-training with progressive representation enhancement for unsupervised cross-domain person re-identification," *IEEE Trans. Image Process.*, vol. 30, pp. 5287–5298, 2021.

[12] K. Wang, P. Wang, C. Ding, and D. Tao, "Batch coherence-driven network for part-aware person re-identification," *IEEE Trans. Image Process.*, vol. 30, pp. 3405–3418, 2021.

[13] S. He et al., "TransReID: Transformer-based object re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 15013–15022.

[14] K. Zhu et al., "AAformer: Auto-aligned transformer for person re-identification," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Aug. 25, 2023, doi: 10.1109/TNNLS.2023.3301856.

[15] H. Ma, X. Li, X. Yuan, and C. Zhao, "Denseformer: A dense transformer framework for person re-identification," *IET Comput. Vis.*, vol. 17, no. 5, pp. 527–536, 2022.

[16] M. Jia, X. Cheng, S. Lu, and J. Zhang, "Learning disentangled representation implicitly via transformer for occluded person re-identification," *IEEE Trans. Multimedia*, vol. 25, pp. 1294–1305, 2023.

[17] L. Shen, T. He, Y. Guo, and G. Ding, "X-ReID: Cross-instance transformer for identity-level person re-identification," 2023, *arXiv:2302.02075*.

[18] Y. Lu, M. Jiang, Z. Liu, and X. Mu, "Dual-branch adaptive attention transformer for occluded person re-identification," *Image Vis. Comput.*, vol. 131, 2023, Art. no. 104633.

[19] C. Chen, M. Ye, M. Qi, J. Wu, J. Jiang, and C.-W. Lin, “Structure-aware positional transformer for visible-infrared person re-identification,” *IEEE Trans. Image Process.*, vol. 31, pp. 2352–2364, 2022.

[20] Y. Li and Y. Chen, “Infrared-visible cross-modal person re-identification via dual-attention collaborative learning,” *Signal Process.: Image Commun.*, vol. 109, 2022, Art. no. 116868.

[21] Y. Feng et al., “Visible-infrared person re-identification via cross-modality interaction transformer,” *IEEE Trans. Multimedia*, vol. 25, pp. 7647–7659, 2023.

[22] K. Jiang, T. Zhang, X. Liu, B. Qian, Y. Zhang, and F. Wu, “Cross-modality transformer for visible-infrared person re-identification,” in *Proc. 17th Eur. Conf. Comput. Vis.*, 2022, pp. 480–496.

[23] J. Zhao, H. Wang, Y. Zhou, R. Yao, S. Chen, and A. El Saddik, “Spatial-channel enhanced transformer for visible-infrared person re-identification,” *IEEE Trans. Multimedia*, vol. 25, pp. 3668–3680, 2023.

[24] K. Wang, H. Wang, M. Liu, X. Xing, and T. Han, “Survey on person re-identification based on deep learning,” *CAAI Trans. Intell. Technol.*, vol. 3, no. 4, pp. 219–227, 2018.

[25] D. Wu et al., “Deep learning-based methods for person re-identification: A comprehensive review,” *Neurocomputing*, vol. 337, pp. 354–371, 2019.

[26] M. O. Almasawa, L. A. Elrefaei, and K. Moria, “A survey on deep learning-based person re-identification systems,” *IEEE Access*, vol. 7, pp. 175228–175247, 2019.

[27] Q. Leng, M. Ye, and Q. Tian, “A survey of open-world person re-identification,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 4, pp. 1092–1108, Apr. 2020.

[28] N. Mathur, S. Mathur, D. Mathur, and P. Dadheech, “A brief survey of deep learning techniques for person re-identification,” in *Proc. IEEE 3rd Int. Conf. Emerg. Technol. Comput. Eng.: Mach. Learn. Internet Things*, 2020, pp. 129–138.

[29] K. Islam, “Person search: New paradigm of person re-identification: A survey and outlook of recent works,” *Image Vis. Comput.*, vol. 101, 2020, Art. no. 103970.

[30] Y. Wang, S. Yang, S. Liu, and Z. Zhang, “Cross-domain person re-identification: A review,” in *Proc. Artif. Intell. China: Proc. 2nd Int. Conf. Artif. Intell.,*2021, pp. 153–160.

[31] C. Yang, F. Qi, and H. Jia, “Survey on unsupervised techniques for person re-identification,” in *Proc. IEEE 2nd Int. Conf. Comput. Data Sci.*, 2021, pp. 161–164.

[32] Y. Peng, S. Hou, C. Cao, X. Liu, Y. Huang, and Z. He, “Deep learning-based occluded person re-identification: A survey,” *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 20, pp. 1–27, 2023.

[33] N. Huang, J. Liu, Y. Miao, Q. Zhang, and J. Han, “Deep learning for visible-infrared cross-modality person re-identification: A comprehensive review,” *Inf. Fusion*, vol. 91, pp. 396–411, 2023.

[34] Z. Ming et al., “Deep learning-based person re-identification methods: A survey and outlook of recent works,” *Image Vis. Comput.*, vol. 119, 2022, Art. no. 104394.

[35] A. Zahra, N. Perwaiz, M. Shahzad, and M. M. Fraz, “Person re-identification: A retrospective on domain specific open challenges and future trends,” *Pattern Recognit.*, vol. 142, 2023, Art. no. 109669.

[36] A. Liberati et al., “The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: Explanation and elaboration,” *Ann. Intern. Med.*, vol. 151, no. 4, pp. W–65, 2009.

[37] S. Xu, J. Wang, W. Shou, T. Ngo, A.-M. Sadick, and X. Wang, “Computer vision techniques in construction: A critical review,” *Arch. Comput. Methods Eng.*, vol. 28, pp. 3383–3397, 2021.

[38] S. Katoch, S. S. Chauhan, and V. Kumar, “A review on genetic algorithm: Past, present, and future,” *Multimedia Tools Appl.*, vol. 80, pp. 8091–8126, 2021.

[39] A. Budrionis, D. Plikynas, P. Daniušis, and A. Indrulionis, “Smartphone-based computer vision travelling aids for blind and visually impaired individuals: A systematic review,” *Assistive Technol.*, vol. 34, no. 2, pp. 178–194, 2022.

[40] D. Gray, S. Brennan, and H. Tao, “Evaluating appearance models for recognition, reacquisition, and tracking,” in *Proc. IEEE Int. Workshop Perform. Eval. Tracking Surveill.*, vol. 3, no. 5, 2007, pp. 1–7.

[41] C. C. Loy, T. Xiang, and S. Gong, “Multi-camera activity correlation analysis,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 1988–1995.

[42] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino, “Custom pictorial structures for re-identification,” in *Proc. Brit. Mach. Vis. Conf.*, 2011, pp. 1–11.

[43] S. Wang, M. Lewandowski, J. Annesley, and J. Orwell, “Re-identification of pedestrians with variable occlusion and scale,” in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2011, pp. 1876–1882.

[44] N. Martinel and C. Micheloni, “Re-identify people in wide area camera network,” in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, 2012, pp. 31–36.

[45] W. Li, R. Zhao, and X. Wang, “Human reidentification with transferred metric learning,” in *Proc. 11th Asian Conf. Comput. Vis. Comput. Vis.*, 2013, pp. 31–44.

[46] W. Li and X. Wang, “Locally aligned feature transforms across views,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 3594–3601.

[47] W. Li, R. Zhao, T. Xiao, and X. Wang, “DeepReID: Deep filter pairing neural network for person re-identification,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 152–159.

[48] A. Das, A. Chakraborty, and A. K. Roy-Chowdhury, “Consistent re-identification in a camera network,” in *Proc. 13th Eur. Conf. Comput. Vis.*, 2014, pp. 330–345.

[49] T. Wang, S. Gong, X. Zhu, and S. Wang, “Person re-identification by discriminative selection in video ranking,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 12, pp. 2501–2514, Dec. 2016.

[50] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, “Scalable person re-identification: A benchmark,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1116–1124.

[51] L. Ma, H. Liu, L. Hu, C. Wang, and Q. Sun, “Orientation driven bag of appearances for person re-identification,” 2016, *arXiv:1605.02464*.

[52] L. Zheng, H. Zhang, S. Sun, M. Chandraker, Y. Yang, and Q. Tian, “Person re-identification in the wild,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1367–1376.

[53] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, “Performance measures and a data set for multi-target, multi-camera tracking,” in *Proc. Comput. Vis. Workshops*, 2016, pp. 17–35.

[54] M. Gou et al., “A systematic evaluation and benchmark for person re-identification: Features, metrics, and datasets,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 3, pp. 523–536, Mar. 2019.

[55] L. Wei, S. Zhang, W. Gao, and Q. Tian, “Person transfer GAN to bridge domain gap for person re-identification,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 79–88.

[56] G. Song, B. Leng, Y. Liu, C. Hetang, and S. Cai, “Region-based quality estimation network for large-scale person re-identification,” in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 7347–7354.

[57] L. Pang, Y. Wang, Y.-Z. Song, T. Huang, and Y. Tian, “Cross-domain adversarial feature learning for sketch re-identification,” in *Proc. 26th ACM Int. Conf. Multimedia*, 2018, pp. 609–617.

[58] A. Wu, W.-S. Zheng, H.-X. Yu, S. Gong, and J. Lai, “RGB-infrared cross-modality person re-identification,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 5380–5389.

[59] D. T. Nguyen, H. G. Hong, K. W. Kim, and K. R. Park, “Person recognition system based on a combination of body images from visible light and thermal cameras,” *Sensors*, vol. 17, no. 3, 2017, Art. no. 605.

[60] Y. Zhang and H. Wang, “Diverse embedding expansion network and low-light cross-modality benchmark for visible-infrared person re-identification,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 2153–2162.

[61] X. Wang, G. Doretto, T. Sebastian, J. Rittscher, and P. Tu, “Shape and appearance context modeling,” in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, 2007, pp. 1–8.

[62] D. Wu, S.-J. Zheng, W.-Z. Bao, X.-P. Zhang, C.-A. Yuan, and D.-S. Huang, “A novel deep model with multi-loss and efficient training for person re-identification,” *Neurocomputing*, vol. 324, pp. 69–75, 2019.

[63] Y. Hao, N. Wang, J. Li, and X. Gao, “HSME: Hypersphere manifold embedding for visible thermal person re-identification,” in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 8385–8392.

[64] M. Ye, X. Lan, J. Li, and P. Yuen, “Hierarchical discriminative learning for visible thermal person re-identification,” in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 7501–7508.

[65] X. Hu and Y. Zhou, “Cross-modality person reid with maximum intra-class triplet loss,” in *Proc. Pattern Recognit. Comput. Vis.: 3rd Chin. Conf.*, 2020, pp. 557–568.

[66] W. Chen, X. Chen, J. Zhang, and K. Huang, “Beyond triplet loss: A deep quadruplet network for person re-identification,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 403–412.

[67] Y. Zhu, Z. Yang, L. Wang, S. Zhao, X. Hu, and D. Tao, “Hetero-center loss for cross-modality person re-identification,” *Neurocomputing*, vol. 386, pp. 97–109, 2020.

[68] X. Fan, W. Jiang, H. Luo, and M. Fei, "Spherereid: Deep hypersphere manifold embedding for person re-identification," *J. Vis. Commun. Image Representation*, vol. 60, pp. 51–58, 2019.

[69] T. Liang et al., "CMTR: Cross-modality transformer for visible-infrared person re-identification," 2021, *arXiv:2110.08994*.

[70] L. Wei, S. Zhang, H. Yao, W. Gao, and Q. Tian, "Glad: Global-local-alignment descriptor for pedestrian retrieval," in *Proc. 25th ACM Int. Conf. Multimedia*, 2017, pp. 420–428.

[71] Y. Shen, T. Xiao, S. Yi, D. Chen, X. Wang, and H. Li, "Person re-identification with deep kronecker-product matching and group-shuffling random walk," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 5, pp. 1649–1665, May 2021.

[72] M. S. Sarfraz, A. Schumann, A. Eberle, and R. Stiefelhagen, "A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 420–429.

[73] L. Zheng, Y. Huang, H. Lu, and Y. Yang, "Pose-invariant embedding for deep person re-identification," *IEEE Trans. Image Process.*, vol. 28, no. 9, pp. 4500–4509, Sep. 2019.

[74] Y.-J. Cho and K.-J. Yoon, "PaMM: Pose-aware multi-shot matching for improving person re-identification," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3739–3752, Aug. 2018.

[75] W. Yang, H. Huang, Z. Zhang, X. Chen, K. Huang, and S. Zhang, "Towards rich feature discovery with class activation maps augmentation for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1389–1398.

[76] J. Xu, R. Zhao, F. Zhu, H. Wang, and W. Ouyang, "Attention-aware compositional network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2119–2128.

[77] Z. Zhang, C. Lan, W. Zeng, X. Jin, and Z. Chen, "Relation-aware global attention for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3186–3195.

[78] S. Gao, J. Wang, H. Lu, and Z. Liu, "Pose-guided visible part matching for occluded person ReID," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11744–11752.

[79] Y.-C. Chen, W.-S. Zheng, J.-H. Lai, and P. C. Yuen, "An asymmetric distance model for cross-view feature mapping in person reidentification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 8, pp. 1661–1675, Aug. 2017.

[80] Z. Wu, Y. Li, and R. J. Radke, "Viewpoint invariant human re-identification in camera networks using pose priors and subject-discriminative features," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 5, pp. 1095–1108, May 2015.

[81] J. Jia, Q. Ruan, G. An, and Y. Jin, "Multiple metric learning with query adaptive weights and multi-task re-weighting for person re-identification," *Comput. Vis. Image Understanding*, vol. 160, pp. 87–99, 2017.

[82] H.-X. Yu, A. Wu, and W.-S. Zheng, "Cross-view asymmetric metric learning for unsupervised person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 994–1002.

[83] H. Chen, Y. Wang, B. Lagadec, A. Dantcheva, and F. Bremond, "Joint generative and contrastive learning for unsupervised person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 2004–2013.

[84] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 480–496.

[85] L. He, Y. Wang, W. Liu, H. Zhao, Z. Sun, and J. Feng, "Foreground-aware pyramid reconstruction for alignment-free occluded person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 8450–8459.

[86] Z. Ma, Y. Zhao, and J. Li, "Pose-guided inter-and intra-part relational transformer for occluded person re-identification," in *Proc. 29th ACM Int. Conf. Multimedia*, 2021, pp. 1487–1496.

[87] M. Jia et al., "Matching on sets: Conquer occluded person re-identification without alignment," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 1673–1681.

[88] Y. Wang, P. Zhang, S. Gao, X. Geng, H. Lu, and D. Wang, "Pyramid spatial-temporal aggregation for video-based person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 12026–12035.

[89] A. Aich, M. Zheng, S. Karanam, T. Chen, A. K. Roy-Chowdhury, and Z. Wu, "Spatio-temporal representation factorization for video-based person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 152–162.

[90] C. Eom, G. Lee, J. Lee, and B. Ham, "Video-based person re-identification with spatial and temporal memory networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 12036–12045.

[91] X. Qian, Y. Fu, Y.-G. Jiang, T. Xiang, and X. Xue, "Multi-scale deep learning architectures for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 5399–5408.

[92] F. Zheng et al., "Pyramidal person re-identification via multi-loss dynamic training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 8514–8522.

[93] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang, "Omni-scale feature learning for person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 3702–3712.

[94] C. Yan et al., "BV-person: A large-scale dataset for bird-view person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 10943–10952.

[95] X. Qian, Y. Fu, T. Xiang, Y.-G. Jiang, and X. Xue, "Leader-based multi-scale attention deep architecture for person re-identification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 371–385, Feb. 2020.

[96] G. Chen, T. Gu, J. Lu, J.-A. Bao, and J. Zhou, "Person re-identification via attention pyramid," *IEEE Trans. Image Process.*, vol. 30, pp. 7663–7676, 2021.

[97] Y. Dai, J. Liu, Y. Bai, Z. Tong, and L.-Y. Duan, "Dual-refinement: Joint label and feature refinement for unsupervised domain adaptive person re-identification," *IEEE Trans. Image Process.*, vol. 30, pp. 7815–7829, 2021.

[98] X. Zhang, Y. Ge, Y. Qiao, and H. Li, "Refining pseudo labels with clustering consensus over generations for unsupervised object re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 3436–3445.

[99] S. Liao and L. Shao, "Interpretable and generalizable person re-identification with query-adaptive convolution and temporal lifting," in *Proc. 16th Eur. Conf. Comput. Vis.*, 2020, pp. 456–474.

[100] J. Liu, Z.-J. Zha, D. Chen, R. Hong, and M. Wang, "Adaptive transfer network for cross-domain person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 7202–7211.

[101] S. Zhou, J. Wang, J. Wang, Y. Gong, and N. Zheng, "Point to set similarity based deep feature learning for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3741–3750.

[102] Y. Tang, X. Yang, N. Wang, B. Song, and X. Gao, "CGAN-TM: A novel domain-to-domain transferring method for person re-identification," *IEEE Trans. Image Process.*, vol. 29, pp. 5641–5651, 2020.

[103] W. Deng, L. Zheng, Q. Ye, G. Kang, Y. Yang, and J. Jiao, "Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 994–1003.

[104] W. Li, X. Zhu, and S. Gong, "Scalable person re-identification by harmonious attention," *Int. J. Comput. Vis.*, vol. 128, no. 6, pp. 1635–1653, 2020.

[105] D. Fu et al., "Improving person re-identification with iterative impression aggregation," *IEEE Trans. Image Process.*, vol. 29, pp. 9559–9571, 2020.

[106] S. Xu, Y. Cheng, K. Gu, Y. Yang, S. Chang, and P. Zhou, "Jointly attentive spatial-temporal pooling networks for video-based person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 4733–4742.

[107] K. Han et al., "A survey on visual transformer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 87–110, Jan. 2023.

[108] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM Comput. Surv.*, vol. 54, no. 10s, pp. 1–41, 2022.

[109] Y. Lu, W. Qin, C. Zhou, and Z. Liu, "Automated detection of dangerous work zone for crawler crane guided by UAV images VIA swin transformer," *Automat. Construction*, vol. 147, 2023, Art. no. 104744.

[110] K. W. Lee, B. Jawade, D. Mohan, S. Setlur, and V. Govindaraju, "Attribute de-biased vision transformer (AD-ViT) for long-term person re-identification," in *Proc. IEEE 18th Int. Conf. Adv. Video Signal Based Surveill.*, 2022, pp. 1–8.

[111] Y. Li, J. He, T. Zhang, X. Liu, Y. Zhang, and F. Wu, "Diverse part discovery: Occluded person re-identification with part-aware transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 2898–2907.

[112] B. Xu, "Region selection for occluded person re-identification via policy gradient," *Image Vis. Comput.*, vol. 132, 2023, Art. no. 104648.

[113] T. Wang, H. Liu, P. Song, T. Guo, and W. Shi, "Pose-guided feature disentangling for occluded person re-identification based on transformer," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 2540–2549.

[114] Y. Zhao, S. Zhu, D. Wang, and Z. Liang, "Short range correlation transformer for occluded person re-identification," *Neural Comput. Appl.*, vol. 34, no. 20, pp. 17633–17645, 2022.

[115] X. Ren, D. Zhang, and X. Bao, "Person re-identification with a cloth-changing aware transformer," in *Proc. Int. Joint Conf. Neural Netw.*, 2022, pp. 1–8.

[116] V. Bansal, G. L. Foresti, and N. Martinel, "Cloth-changing person re-identification with self-attention," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2022, pp. 602–610.

[117] X. Liu, P. Zhang, C. Yu, H. Lu, X. Qian, and X. Yang, "A video is worth three views: Trigeminal transformers for video-based person re-identification," 2021, *arXiv:2104.01745*.

[118] T. Zhang et al., "Spatiotemporal transformer for video-based person re-identification," 2021, *arXiv:2103.16469*.

[119] H. Yao, S. Zhang, R. Hong, Y. Zhang, C. Xu, and Q. Tian, "Deep representation learning with part loss for person re-identification," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2860–2871, Jun. 2019.

[120] H. Li, M. Ye, C. Wang, and B. Du, "Pyramidal transformer with conv-patchify for person re-identification," in *Proc. 30th ACM Int. Conf. Multimedia*, 2022, pp. 7317–7326.

[121] K. Li et al., "UniFormer: Unifying convolution and self-attention for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 10, pp. 12581–12600, Oct. 2023.

[122] B. Tan, L. Xu, Z. Qiu, Q. Wu, and F. Meng, "MFAT: A multi-level feature aggregated transformer for person re-identification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2023, pp. 1–5.

[123] Z. Liu, X. Mu, Y. Lu, T. Zhang, and Y. Tian, "Learning transformer-based attention region with multiple scales for occluded person re-identification," *Comput. Vis. Image Understanding*, vol. 229, 2023, Art. no. 103652.

[124] F. Yang, W. Li, B. Liang, and J. Zhang, "Spatiotemporal interaction transformer network for video-based person re-identification in Internet of Things," *IEEE Internet Things J.*, vol. 10, no. 14, pp. 12537–12547, Jul. 2023.

[125] X. Liu, C. Yu, P. Zhang, and H. Lu, "Deeply coupled convolution–transformer with spatial–temporal complementary learning for video-based person re-identification," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, May 26, 2023, doi: 10.1109/TNNLS.2023.3271353.

[126] T. Liang, Y. Jin, W. Liu, and Y. Li, "Cross-modality transformer with modality mining for visible-infrared person re-identification," *IEEE Trans. Multimedia*, vol. 25, pp. 8432–8444, 2023.

[127] Z. Chai, Y. Ling, Z. Luo, D. Lin, M. Jiang, and S. Li, "Dual-stream transformer with distribution alignment for visible-infrared person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 11, pp. 6764–6776, Nov. 2023.

[128] K. Han, A. Xiao, E. Wu, J. Guo, C. Xu, and Y. Wang, "Transformer in transformer," in*Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 15908–15919.

[129] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 10012–10022.

[130] C.-F. Chen, R. Panda, and Q. Fan, "RegionViT: Regional-to-local attention for vision transformers," 2021, *arXiv:2106.02689*.

[131] Z. Huang, Y. Ben, G. Luo, P. Cheng, G. Yu, and B. Fu, "Shuffle transformer: Rethinking spatial shuffle for vision transformer," 2021, *arXiv:2106.03650*.

[132] J. Fang, L. Xie, X. Wang, X. Zhang, W. Liu, and Q. Tian, "MSG-transformer: Exchanging local spatial information by manipulating messenger tokens," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 12063–12072.

[133] L. Yuan et al., "Tokens-to-token ViT: Training vision transformers from scratch on imagenet," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 558–567.

[134] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[135] K. Han et al., "A survey on vision transformer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 87–110, Jan. 2023.

[136] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 10347–10357.

[137] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "ViViT: A video vision transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 6836–6846.

[138] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. 16th Eur. Conf. Comput. Vis.*, 2020, pp. 213–229.

[139] S. Zheng et al., "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 6881–6890.

[140] C. Sharma, S. R. Kapil, and D. Chapman, "Person re-identification with a locally aware transformer," 2021, *arXiv:2106.03720*.

[141] Q. Wang et al., "Swin transformer based on two-fold loss and background adaptation re-ranking for person re-identification," *Electronics*, vol. 11, no. 13, 2022, Art. no. 1941.

[142] B. Zhang, Y. Liang, and M. Du, "Interlaced perception for person re-identification based on swin transformer," in *Proc. IEEE 7th Int. Conf. Image, Vis. Comput.*, 2022, pp. 24–30.

[143] G. Zhang, P. Zhang, J. Qi, and H. Lu, "HAT: Hierarchical aggregation transformers for person re-identification," in *Proc. 29th ACM Int. Conf. Multimedia*, 2021, pp. 516–525.

[144] P. Li, P. Pan, P. Liu, M. Xu, and Y. Yang, "Hierarchical temporal modeling with mutual distance matching for video based person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 2, pp. 503–511, Feb. 2021.

[145] J. Li, M. Wang, and X. Gong, "Transformer based multi-grained features for unsupervised person re-identification," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2023, pp. 42–50.

[146] S. Lai, Z. Chai, and X. Wei, "Transformer meets part model: Adaptive part division for person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 4150–4157.

[147] H. Luo et al., "Self-supervised pre-training for transformer-based person re-identification," 2021, *arXiv:2111.12084*.

[148] Y. Chen et al., "ResT-ReID: Transformer block-based residual learning for person re-identification," *Pattern Recognit. Lett.*, vol. 157, pp. 90–96, 2022.

[149] Z. Wang, X. Jiang, K. Xu, and T. Sun, "A transformer-based cloth-irrelevant patches feature extracting method for long-term cloth-changing person re-identification," in *Proc. Adv. Comput. Graph.: 39th Comput. Graph. Int. Conf.*, 2023, pp. 278–289.

[150] G. Wang, T. Zhang, J. Cheng, S. Liu, Y. Yang, and Z. Hou, "RGB-infrared cross-modality person re-identification via joint pixel and feature alignment," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 3623–3632.

[151] W. Luo, Y. Li, R. Urtasun, and R. Zemel, "Understanding the effective receptive field in deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 4898–4906.

[152] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang, "Bag of tricks and a strong baseline for deep person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2019, pp. 1487–1495.

[153] X. Jin, C. Lan, W. Zeng, G. Wei, and Z. Chen, "Semantics-aligned representation learning for person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 07, pp. 11173–11180.

[154] X. Chen et al., "Salience-guided cascaded suppression network for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3300–3310.

[155] K. Zhu, H. Guo, Z. Liu, M. Tang, and J. Wang, "Identity-guided human semantic parsing for person re-identification," in *Proc. 16th Eur. Conf. Comput. Vis.*, 2020, pp. 346–363.

[156] Z. Zhang, H. Zhang, and S. Liu, "Person re-identification using heterogeneous local graph attention networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 12136–12145.

[157] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[158] G.-A. Wang et al., "Cross-modality paired-images generation for rgb-infrared person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 07, pp. 12144–12151.

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

[160] A. Wu, W.-S. Zheng, S. Gong, and J. Lai, "RGB-IR person re-identification by cross-modality similarity preservation," *Int. J. Comput. Vis.*, vol. 128, pp. 1765–1785, 2020.

[161] M. Ye, J. Shen, D. J. Crandall, L. Shao, and J. Luo, "Dynamic dual-attentive aggregation learning for visible-infrared person re-identification," in *Proc. 16th Eur. Conf. Comput. Vis.*, 2020, pp. 229–247.

[162] Y. Lu et al., "Cross-modality person re-identification with shared-specific feature transfer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 13379–13389.

[163] Y. Chen, L. Wan, Z. Li, Q. Jing, and Z. Sun, "Neural feature search for RGB-infrared person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 587–597.

[164] Z. Zhao, B. Liu, Q. Chu, Y. Lu, and N. Yu, "Joint color-irrelevant consistency learning and identity-aware modality adaptation for visible-infrared cross modality person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 3520–3528.

[165] D. Zhang, Z. Zhang, Y. Ju, C. Wang, Y. Xie, and Y. Qu, "Dual mutual learning for cross-modality person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 8, pp. 5361–5373, Aug. 2022.

[166] M. Ye, J. Shen, and L. Shao, "Visible-infrared person re-identification via homogeneous augmented tri-modal learning," *IEEE Trans. Inf. Forensics Secur.*, vol. 16, pp. 728–739, 2021.

[167] Y. Huang, Q. Wu, J. Xu, Y. Zhong, P. Zhang, and Z. Zhang, "Alleviating modality bias training for infrared-visible person re-identification," *IEEE Trans. Multimedia*, vol. 24, pp. 1570–1582, 2022.

[168] J. Sun, Y. Li, H. Chen, Y. Peng, X. Zhu, and J. Zhu, "Visible-infrared cross-modality person re-identification based on whole-individual training," *Neurocomputing*, vol. 440, pp. 1–11, 2021.

[169] K. Li, X. Wang, Y. Liu, B. Zhang, and M. Zhang, "Cross-modality disentanglement and shared feedback learning for infrared-visible person re-identification," *Knowl.-Based Syst.*, vol. 252, 2022, Art. no. 109337.

[170] M. Ye, C. Chen, J. Shen, and L. Shao, "Dynamic tri-level relation mining with attentive graph for visible infrared re-identification," *IEEE Trans. Inf. Forensics Secur.*, vol. 17, pp. 386–398, 2022.

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

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

[173] M. Yang, Z. Huang, P. Hu, T. Li, J. Lv, and X. Peng, "Learning with twin noisy labels for visible-infrared person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 14308–14317.

[174] J. Liu, J. Liu, and Q. Zhang, "M2FINet: Modality-specific and modality-shared features interaction network for RGB-IR person re-identification," *Comput. Vis. Image Understanding*, vol. 232, 2023, Art. no. 103708.

**Prodip Kumar Sarker** received the M.Sc. degree in computer science and engineering from the University of Rajshahi, Rajshahi, Bangladesh. He is currently working toward the Ph.D. degree with the School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China. His research interests include computer vision, deep learning, transformer, and person re-identification.

**Qingjie Zhao** received the Ph.D. degree in computer science and technology from Tsinghua University, Beijing, China, in 2003. She is currently with the School of Computer Science and Technology, Beijing Institute of Technology, Beijing. From 2008 to 2009, she was a visiting Fellow with the School of Computer Science and Electronic Engineering, University of Essex, Colchester, U.K. From 2017 to 2018, she was a Senior Visiting Scholar with the Department of Information, University of Hamburg, Hamburg, Germany. Her research interests include image and video processing, machine learning, and intelligent system. She is a Member of China Computer Federation, Chinese Association of Automation, and China Association of Artificial Intelligence.

**Md. Kamal Uddin** received the Ph.D. degree in information and computer sciences from Saitama University, Saitama, Japan, in 2021. He is currently an Associate Professor with the Department of Computer Science and Telecommunication Engineering, Noakhali Science and Technology University, Noakhali, Bangladesh. His research interests include computer vision, machine learning techniques for image and video processing, person re-identification, and video surveillance techniques.