DOI: 10.65322/j.srd.04182026.24

LOW-RESOURCE SPATIAL-SEMANTIC TRANSFORMER FEATURE EXTRACTION FOR PERSON RE-IDENTIFICATION IN CROWDED VIDEO SURVEILLANCE

Begmatov Shohruh, *Doctoral Researcher,*
Arabboev Mukhriddin, *Doctoral Researcher,*
Nishanov Akhram, *Professor,*
Tashkent University of Information Technologies named after Muhammad al-Khwarizmi

Abstract: *Person re-identification in crowded video surveillance is difficult due to occlusion, scale variation, viewpoint changes, and visually similar pedestrians. Existing deep models often provide strong accuracy but require high computational resources. This paper aims to develop a low-resource spatial-semantic feature extractor that produces compact and discriminative identity embeddings for real-time person re-identification. The proposed model processes each 3 × 192 × 64 full-body pedestrian crop using a 6 × 2 spatial grid and a semantic global branch. Local spatial tokens and global semantic features are fused, enhanced with positional encoding, and refined by lightweight transformer blocks. The final output is a 384-dimensional L2-normalized embedding trained using cosine-triplet learning. Experiments use a MOT20-derived dataset containing 387,426 pedestrian crops, 2,089 identities, and 310,602 triplets. The proposed model provides effective separation between same-identity and different-identity samples while maintaining compact representation and low computational cost. Similarity, distance, and threshold-based evaluations show that the spatial-semantic token design improves identity matching in crowded scenes. The proposed feature extractor offers an efficient solution for person re-identification in resource-constrained surveillance systems by combining local body-region information, global appearance features, and lightweight transformer refinement.*

Annotatsiya: *Tiqilinch videokuzatuv sharoitlarida shaxsni qayta identifikatsiyalash okklyuziya, masshtab o‘zgarishi, kuzatish burchagining o‘zgarishi va tashqi ko‘rinishi o‘xshash piyodalar mavjudligi sababli murakkab vazifa hisoblanadi. Mavjud chuqur o‘rganish modellari ko‘p hollarda yuqori aniqlikni ta’minlaydi, biroq katta hisoblash resurslarini talab qiladi. Ushbu maqolaning maqsadi real vaqt rejimida shaxsni qayta identifikatsiyalash uchun ixcham va ajratuvchan identifikatsion embeddinglarni hosil qiluvchi kam resurs talab qiladigan fazoviy-semantik belgilar ajratgichini ishlab chiqishdan iborat. Taklif etilgan model har bir 3 × 192 × 64 o‘lchamli, piyodaning butun tanasi aks etgan kesmani 6 × 2 fazoviy to‘r va semantik global tarmoq yordamida qayta ishlaydi. Lokal fazoviy tokenlar va global semantik belgilar birlashtiriladi, pozitsion kodlash bilan boyitiladi va yengil transformer bloklari orqali takomillashtiriladi. Yakuniy natija kosinus-tripletli o‘qitish asosida o‘rgatilgan, L2-normallashtirilgan 384 o‘lchamli embeddingdan iborat. Tajribalar MOT20 asosida hosil qilingan 387 426 ta piyoda kesmasi, 2 089 ta identifikator va 310 602 ta tripletni o‘z ichiga olgan ma’lumotlar to‘plamida o‘tkazildi. Taklif etilgan model ifoda ixchamligi va past hisoblash sarfini saqlagan holda bir shaxsga tegishli va turli shaxslarga tegishli namunalarni samarali ajratadi. O‘xshashlik, masofa va porog qiymatlari asosidagi baholashlar fazoviy-semantik tokenlar dizayni tiqilinch sahnalarda shaxslarni moslash aniqligini yaxshilashini ko‘rsatdi. Taklif etilgan belgilar ajratgichi lokal tana sohalari haqidagi ma’lumot, global tashqi ko‘rinish belgilari va yengil transformerli takomillashtirishni birlashtirish orqali resurslari cheklangan videokuzatuv tizimlarida shaxsni qayta identifikatsiyalash uchun samarali yechimni taklif etadi.*

Аннотация: *Реидентификация людей в многолюдных сценах видеонаблюдения является сложной задачей из-за окклюзии, изменения масштаба, ракурса и внешне похожих пешеходов. Существующие модели глубокого обучения часто обеспечивают*

DOI: 10.65322/j.srd.04182026.24

высокую точность, но требуют значительных вычислительных ресурсов. Цель данной работы — разработать малоресурсный пространственно-семантический экстрактор признаков, формирующий компактные и дискриминативные эмбеддинги идентичности для реидентификации людей в реальном времени. Предлагаемая модель обрабатывает каждый фрагмент изображения пешехода в полный рост размером 3 × 192 × 64 с использованием пространственной сетки 6 × 2 и семантической глобальной ветви. Локальные пространственные токены и глобальные семантические признаки объединяются, дополняются позиционным кодированием и уточняются легковесными трансформерными блоками. На выходе формируется 384-мерный L2-нормализованный эмбеддинг, обучаемый с использованием косинусной триплетной функции потерь. Эксперименты проведены на наборе данных, полученном на основе MOT20 и содержащем 387 426 фрагментов изображений пешеходов, 2 089 идентичностей и 310 602 триплета. Предлагаемая модель обеспечивает эффективное разделение образцов одной и разных идентичностей при сохранении компактности представления и низких вычислительных затрат. Оценка на основе сходства, расстояния и пороговых значений показывает, что пространственно-семантическая организация токенов улучшает сопоставление идентичностей в многолюдных сценах. Предлагаемый экстрактор признаков представляет собой эффективное решение для реидентификации людей в системах видеонаблюдения с ограниченными ресурсами за счет объединения информации о локальных областях тела, глобальных признаков внешности и легковесного трансформерного уточнения.

Keywords: *Person re-identification; spatial-semantic feature extraction; transformer refinement; metric learning; triplet loss; crowded video surveillance; MOT20; lightweight deep learning.*

Kalit so‘zlar: *Shaxsni qayta identifikatsiyalash; fazoviy-semantik belgilarni ajratib olish; transformerli takomillashtirish; metrik o‘rganish; triplet yo‘qotish funksiyasi; tiqilinch videokuzatuv; MOT20; yengil chuqur o‘rganish.*

Ключевые слова: *Реидентификация людей; пространственно-семантическое извлечение признаков; трансформерное уточнение; метрическое обучение; триплетная функция потерь; видеонаблюдение в многолюдных сценах; MOT20; легковесное глубокое обучение.*

1. Introduction

Person re-identification is an essential task in intelligent video surveillance, multi-object tracking, and smart city monitoring. Its purpose is to determine whether pedestrian images captured at different frames, viewpoints, or camera locations belong to the same identity. In crowded scenes, this task becomes difficult because of occlusion, scale variation, pose changes, motion blur, and visually similar clothing.

Deep learning has significantly improved person re-identification through convolutional feature extractors, metric learning, and transformer-based representation models. However, many high-performing methods depend on large backbones, high-dimensional embeddings, and computationally expensive architectures. This limits their practical use in real-time surveillance systems, where many pedestrian crops must be processed repeatedly. Similar efficiency concerns have also been reported in lightweight visual detection studies, where reducing model complexity is considered important for real-time intelligent vision systems [1].

To address this limitation, this paper proposes a low-resource spatial-semantic transformer feature extractor for person re-identification. The proposed model processes a full-body pedestrian crop of size 3 × 192 × 64 and divides it into a 6 × 2 grid, generating 12 local spatial tokens. A semantic branch extracts a global feature from the whole crop. Local and global features are then

DOI: 10.65322/j.srd.04182026.24

fused, enhanced with two-dimensional positional encoding, and refined using lightweight transformer blocks. The final output is a 384-dimensional L2-normalized identity embedding used for cosine-similarity-based matching.

The method is evaluated using a MOT20-derived full-body pedestrian crop and triplet dataset. The dataset contains 387,426 curated crop images, 2,089 identities, and 310,602 triplet annotations, organized into train, validation, query, and gallery splits. This structure supports both metric-learning-based training and retrieval-style person re-identification evaluation.

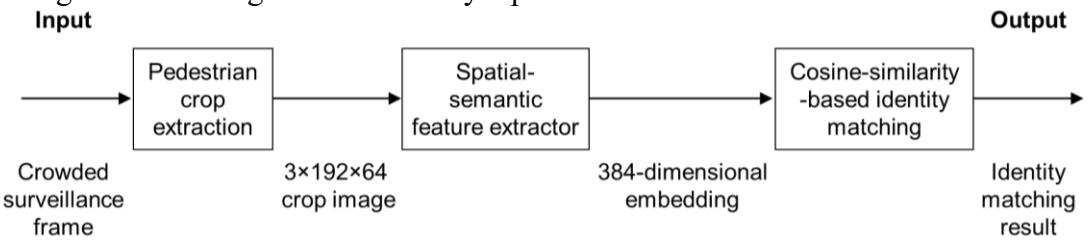


Figure 1. Overall workflow of the proposed low-resource person re-identification framework

The key contributions of this work are summarized as follows:

– A low-resource spatial-semantic feature extractor is proposed for person re-identification in crowded video surveillance;

– A compact 6 × 2 spatial token representation is introduced to preserve body-region information with reduced computational cost;

– Local spatial features and global semantic appearance are fused in a unified token-based representation;

– The proposed model is evaluated using a MOT20-derived full-body crop and triplet dataset designed for metric-learning-based person re-identification;

The rest of this paper is organized as follows. Section 2 reviews related work. Section 3 describes the dataset and proposed method. Section 4 presents the experimental setup. Section 5 discusses the results, and Section 6 concludes the paper.

2. Literature Review

Person re-identification has become an important research direction in computer vision because of its direct application in surveillance, multi-camera monitoring, pedestrian retrieval, and intelligent tracking systems. The main goal of person re-identification is to learn a discriminative representation that can match the same person across different frames, viewpoints, cameras, and environmental conditions. Recent studies show that this task remains challenging because of occlusion, pose variation, illumination changes, low image quality, background clutter, and visual similarity between different pedestrians [2], [3].

In related intelligent-vision applications, object detection has also been widely studied as a fundamental perception task, including deep-learning-based, salient-object, three-dimensional, and end-to-end detection approaches [4].

2.1. CNN-Based Person Re-Identification

Convolutional neural networks have been widely used as backbone models for person re-identification. Residual networks such as ResNet50, ResNet101, and ResNet152 are frequently adopted because residual connections allow deeper feature extraction while reducing the degradation problem in very deep networks [5]. Several person re-identification studies have used ResNet-based architectures as baseline feature extractors or as the main backbone for improved identity representation. For example, ResNet50 has been used in local-global attention, edge-enhanced feature extraction, and CLIP-based person re-identification studies[6]-[9]. ResNet101

DOI: 10.65322/j.srd.04182026.24

and ResNet152 have also been investigated in deeper feature-learning settings, where they provide stronger representational capacity but increase computational cost [10].

VGG16 is another classical convolutional backbone used in visual recognition and person re-identification studies. Its sequential convolutional structure provides a simple and interpretable baseline for deep feature extraction [11]. However, VGG-based models usually contain a large number of parameters and are computationally expensive compared with more modern compact architectures. Therefore, VGG16 is useful mainly as a conventional CNN baseline for comparison with more efficient person re-identification models.

Although CNN-based models are effective in extracting appearance features, global CNN representations may become less reliable in crowded scenes. When pedestrian crops are partially occluded or visually similar, a single global descriptor may fail to preserve important local body-region cues. This limitation has motivated the development of local feature learning, part-based representation, and attention-based methods.

2.2. Local and Global Feature Fusion

Local-global feature fusion has become a major direction in person re-identification. These methods attempt to combine whole-body appearance information with local body-region details. Recent studies on local and global attention mechanisms show that combining different pooling strategies and attention modules can improve feature representation by capturing both fine-grained and holistic pedestrian characteristics [12]. Similarly, edge-enhanced feature extraction and inter-part relationship modeling have been proposed to strengthen the representation of local body structures and improve identity discrimination in complex scenes [13].

The importance of local and global representation is also emphasized in transformer-based person re-identification. A local-global transformer-based model was proposed to improve the interaction between local regions and global representations [14]. Such approaches confirm that identity representation benefits from combining region-level details with whole-image semantic information. However, many local-global models require additional attention modules, multi-branch structures, or larger transformer backbones, which may increase inference time.

The proposed method follows the same general motivation but uses a simpler and lower-resource structure. Instead of using many patches or complex part-alignment modules, the pedestrian crop is divided into a fixed 6 × 2 grid, producing only 12 spatial tokens. These local tokens are fused with a global semantic feature, allowing the model to preserve local body information while maintaining computational efficiency.

2.3. Lightweight Person Re-Identification

Real-time person re-identification requires a balance between accuracy and efficiency. In practical surveillance systems, the feature extractor must process many pedestrian crops repeatedly, and therefore model size, inference time, and memory consumption become important factors. Lightweight person re-identification methods have been proposed to address this issue by using compact networks, multi-attribute feature generation, efficient convolutional blocks, or simplified attention modules [15].

Recent efficient person re-identification studies also explore lightweight transformer structures. TE-TransReID introduces local feature embedding and a lightweight transformer design to reduce computational complexity while maintaining discriminative feature learning [16]. These studies demonstrate that efficient architecture design is necessary for practical deployment. However, even lightweight transformer models may still depend on relatively complex tokenization and feature-processing strategies.

The proposed model is designed as a low-resource alternative. It uses only 12 spatial-semantic tokens and outputs a compact 384-dimensional normalized embedding. This design reduces the number of tokens processed by the transformer block and makes the model suitable for resource-constrained person re-identification tasks.

DOI: 10.65322/j.srd.04182026.24

2.4. Transformer-Based Person Re-Identification

Vision transformers have significantly influenced person re-identification research because self-attention can model long-range dependencies among image regions. The original Vision Transformer showed that images can be represented as sequences of patches and processed using transformer encoders [17]. In person re-identification, transformer-based methods have been used to model body-region relationships, global-local interactions, and cross-modal identity representation.

Recent transformer-based works include PersonViT, which investigates large-scale self-supervised vision transformer pretraining for person re-identification [6], and Completed Part Transformer, which focuses on part interaction and part-global dependencies [18]. CLIP-based and image-text person re-identification methods also use transformer representations to connect visual identity features with textual or multimodal information [19], [20], [21]. These studies indicate that transformer architectures can improve representation learning, especially when local regions must interact with global identity cues.

However, standard vision transformer models such as ViT-B/16, ViT-L/16, and ViT-H/14 are computationally expensive because they process many image patches and contain large numbers of parameters. This limits their direct use in real-time crowded surveillance systems. The proposed method adopts the token-interaction advantage of transformers but reduces complexity by using a small number of fixed spatial tokens and a compact embedding dimension.

2.5. Occlusion and Crowded-Scene Re-Identification

Occlusion is one of the main difficulties in person re-identification. Similar challenges are also observed in robust object detection, where occlusion, scale variation, and complex backgrounds can reduce detection reliability in practical computer-vision systems [22]. In crowded surveillance scenes, pedestrians are often partially hidden by other people or objects, which reduces the reliability of global appearance descriptors. Recent occluded person re-identification studies attempt to recover corrupted features, identify visible body regions, or reduce the effect of missing parts [23]. Vision transformer-based occluded person re-identification methods have also been explored for complex surveillance scenes where viewpoint distortion and poor image quality are common [24].

These studies show that robust identity representation should not rely only on full global appearance. Local body-region information is important because some visible regions may remain discriminative even when other regions are occluded. The proposed spatial-semantic representation addresses this problem by preserving fixed local body-region tokens and combining them with global semantic context.

2.6. Metric Learning and Contrastive Representation

Metric learning is commonly used in person re-identification to learn an embedding space in which same-identity samples are close and different-identity samples are separated. Recent unsupervised and contrastive learning studies show that contrastive objectives, attention mechanisms, and hard sample mining can improve the discriminative ability of person re-identification features [25]. CLIP-based person re-identification methods also demonstrate that normalized embeddings and similarity-based matching are effective for retrieval-style identity comparison [19], [21].

In this paper, the proposed feature extractor produces a 384-dimensional L2-normalized embedding. Identity comparison is performed using cosine similarity and cosine distance, while triplet-loss-based learning is used to improve separation between anchor-positive and anchor-negative samples. This is consistent with the general direction of metric-learning-based person re-identification, but the proposed architecture focuses on reducing computational cost through compact spatial-semantic token modeling.

Table 1

DOI: 10.65322/j.srd.04182026.24

Summary of related ReID studies

Reference	Research direction	Model	Main idea	Limitation
[5]	Vision transformer foundation	ViT-B/16, ViT-L/16, ViT-H/14	Introduced image patch tokenization and transformer-based visual recognition	General vision model; not designed specifically for ReID
[14]	Transformer-based ReID	ViT-based TransReID	Applied transformer architecture to object/person re-identification	Large patch-token transformer is computationally expensive
[6]	Local-global attention ReID	ResNet50, MGN, attention, combination pooling	Combines global and local features with attention and optimized pooling	Multi-branch structure increases complexity
[24]	Part-aware transformer ReID	ViT-B/16-based CPT	Models intra-part and part-global interactions for ReID	More complex than compact 12-token representation
[26]	CLIP-based video ReID	CLIP visual encoder, ViT-B/16	Introduces text-free CLIP adaptation for video-based ReID	Video-oriented and CLIP-dependent design is heavier
[7]	CLIP domain-gap ReID	CLIP ResNet50, CLIP ViT-B/16	Quantitatively studies cross-dataset domain gap in CLIP-based ReID	Focuses on domain-gap analysis, not lightweight model design
[8]	Dual descriptor ReID	ConvNeXt V2, ResNet50/101/152 and ViT-Base	Uses two sub-networks to generate complementary descriptors	Dual-network design increases model size
[9]	CLIP-based image ReID	CLIP ResNet50, CLIP ViT-B/16	Uses learnable ID-specific text tokens for image ReID	Requires CLIP pretraining and two-stage training
[27]	Edge-enhanced part ReID	ResNet-based CNN with edge and part-relation modules	Combines edge features, global features, and inter-part relationships	Additional edge and part modules increase complexity
[28]	Crowded-scene benchmark	MOT20	Provides dense crowded-scene video data used as the source benchmark	Tracking benchmark; requires conversion into ReID crops

2.7. Research Gap

The reviewed studies show that modern person re-identification methods have achieved significant progress through CNN backbones, local-global feature fusion, transformer architectures, and metric learning. However, several limitations remain. First, deep CNN

DOI: 10.65322/j.srd.04182026.24

backbones such as ResNet101, ResNet152, VGG16, and large ViT variants can be computationally expensive. Second, global feature descriptors may be insufficient in crowded and occluded scenes. Third, many local-global or transformer-based models improve accuracy but introduce additional branches, large token sets, or complex attention mechanisms. To address these limitations, this paper proposes a low-resource spatial-semantic transformer feature extractor.

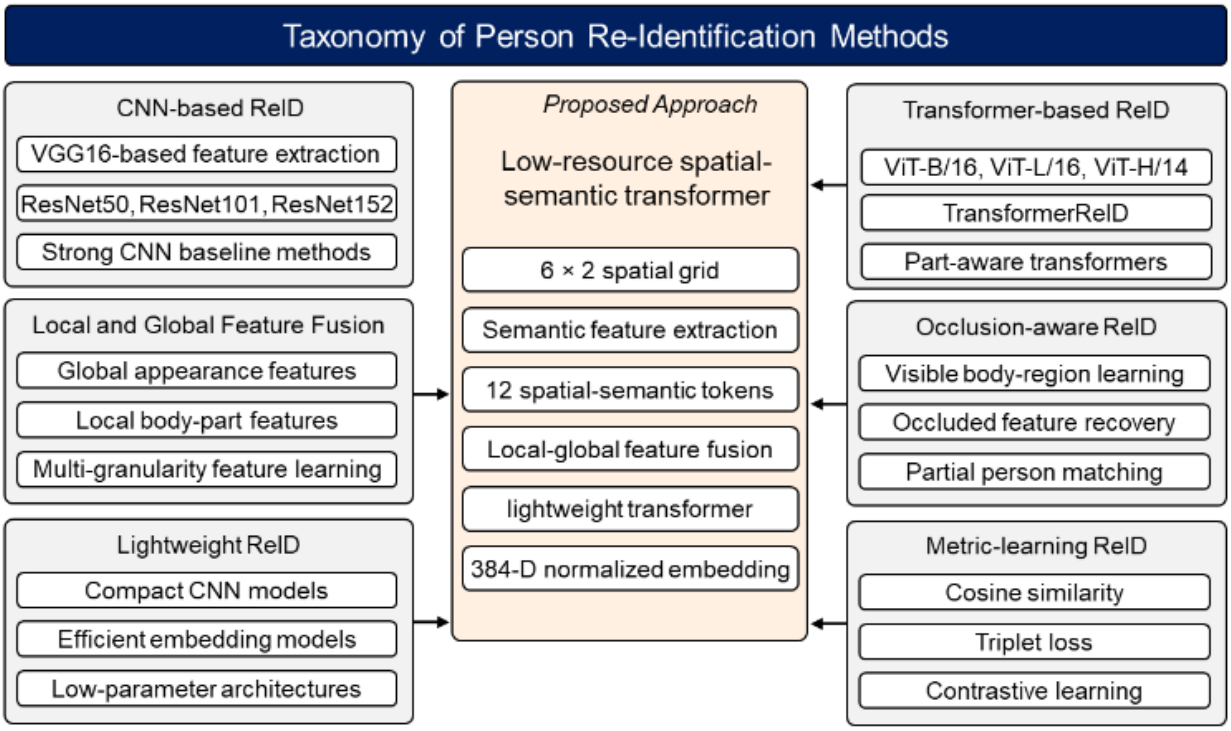


Figure 2. Taxonomy of related person re-identification methods and the position of the proposed low-resource spatial-semantic transformer approach

As shown in Figure 2, the proposed approach integrates local–global feature fusion, lightweight representation learning, transformer-based token interaction, occlusion-aware feature modeling, and metric-learning-based optimization in a compact spatial-semantic architecture.

Table 2

Baseline model usage in related works

Model	Related works	Purpose in prior studies	Limitation for the present study
VGG16	Classical CNN baseline; mentioned as a common pre-trained ReID foundation in edge-enhanced ReID studies	Conventional CNN feature extraction baseline	Large parameter count and lower efficiency
ResNet50	[6],[7],[8],[9],[27]	Strong CNN backbone for local-global, CLIP-based, descriptor-enhanced, and edge-enhanced ReID	Heavier than the proposed compact 12-token model

DOI: 10.65322/j.srd.04182026.24

ResNet101	[8]	Deeper residual CNN baseline for feature comparison	Higher inference cost than ResNet50
ResNet152	[8]	Very deep CNN baseline for high-capacity feature extraction	Computationally expensive for real-time surveillance
ViT-B/16	[5], [24], [26],[7],[9]	Transformer backbone for patch-token ReID, CLIP-ReID, and part-aware modeling	Processes many patch tokens; higher memory and computation
ViT-L/16	[5]	Large transformer baseline for visual representation comparison	High parameter count and deployment cost
ViT-H/14	[5]	High-capacity transformer baseline / large VLM backbone	Too large for low-resource repeated crop-level ReID
Proposed model	This study	Low-resource spatial-semantic ReID feature extractor	Needs broader validation on additional ReID datasets

The method uses a fixed 6 × 2 spatial token structure, fuses local body-region tokens with global semantic appearance, applies lightweight transformer refinement, and generates a compact 384-dimensional normalized embedding. This design aims to preserve discriminative identity information while reducing computational complexity for person re-identification in crowded video surveillance.

3. Materials and Methods

This section describes the dataset, preprocessing procedure, proposed low-resource spatial-semantic feature extractor, embedding generation, similarity computation, and metric-learning objective used for person re-identification in crowded video surveillance.

3.1. Dataset Description

The experiments were conducted using a derived full-body pedestrian crop and triplet dataset constructed from four MOT20 crowded-scene training sequences: MOT20-01, MOT20-02, MOT20-03, and MOT20-05. The dataset was prepared for person re-identification, metric learning, and retrieval-based identity matching. Unlike the original MOT20 benchmark, which is mainly organized for multi-object tracking, the derived dataset provides cropped full-body pedestrian images, identity labels, split metadata, and triplet annotations suitable for embedding-based re-identification experiments.

Table 3

Sequence-wise filtering statistics

Sequence	Original GT annotations	After confidence filtering	After visibility filtering	Final exported crops
MOT20-01	26,647	19,870	9,050	7,238
MOT20-02	202,215	154,742	59,141	49,418
MOT20-03	356,728	313,658	213,706	126,022
MOT20-05	751,330	646,344	335,758	204,748
Total	**1,336,920**	**1,134,614**	**617,655**	**387,426**

DOI: 10.65322/j.srd.04182026.24

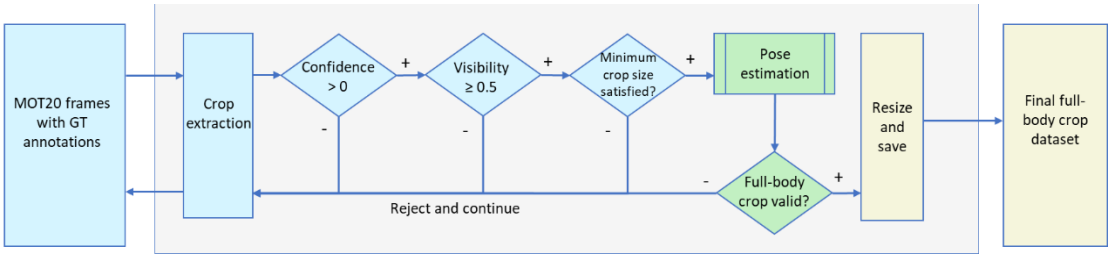


Figure 3. Dataset construction pipeline from MOT20 annotations to full-body ReID crops and triplet annotations

The sequence-wise distribution consists of 7,238 crops from MOT20-01, 49,418 crops from MOT20-02, 126,022 crops from MOT20-03, and 204,748 crops from MOT20-05. In addition to the image crops, the dataset contains metadata files for train, validation, query, and gallery splits, identity mapping files, export statistics, and triplet annotation files in JSONL format. The triplet component contains 310,602 triplets, including 277,688 training triplets, 32,466 validation triplets, and 448 test triplets. Approximately 80% of the triplets are hard-negative samples, enabling the evaluation of visually similar identity-matching cases.

During dataset construction, 1,336,920 original ground-truth pedestrian annotations were examined. After confidence filtering and visibility thresholding with a visibility threshold of 0.5, 617,655 candidate instances remained. Further crop-size filtering and pose-based full-body validation reduced the final exported set to 387,426 full-body pedestrian crop images. The final dataset includes 310,935 training images, 36,603 validation images, 465 query images, and 39,423 gallery images, covering 2,089 exported identities.

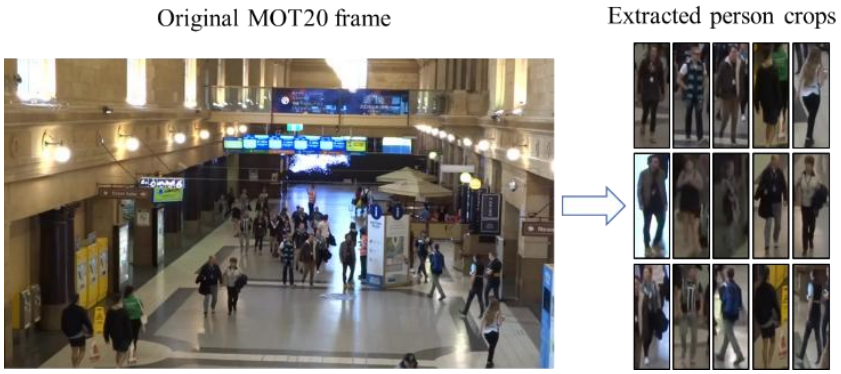


Figure 4. Example of pedestrian crop extraction from crowded MOT20 video frames

All accepted pedestrian crops were resized to 192 × 64 pixels and used as the input format for the proposed model. This resolution is suitable for person re-identification because it preserves the vertical body structure while keeping the computational cost low.

The dataset was divided into training, validation, query, and gallery subsets to support both metric-learning-based training and retrieval-style person re-identification evaluation. The split distribution is shown in Table 4.

Table 4

Dataset split distribution used for person re-identification experiments

Split	Number of images	Number of identities

DOI: 10.65322/j.srd.04182026.24

Training	310,935	2,089
Validation	36,603	2,089
Query	465	465 query samples
Gallery	39,423	Gallery identities from MOT20-derived set
Total	387,426	2,089

As shown in Table 4, the training subset was used to optimize the proposed feature extractor through metric-learning-based training. The validation subset was used to monitor the learning process, evaluate intermediate model performance, and support hyperparameter selection. The query subset was used as the set of probe images in retrieval-style person re-identification evaluation, while the gallery subset served as the search database against which query images were compared. Therefore, the split structure supports both training of the embedding model and evaluation of identity matching under a query-to-gallery retrieval protocol.

3.2. Proposed Low-Resource Spatial-Semantic Feature Extractor

The proposed model is designed to convert each full-body pedestrian crop into a compact and discriminative identity embedding. The model has two main branches: a spatial branch and a semantic branch. The spatial branch preserves local body-region information, while the semantic branch extracts global appearance information from the entire pedestrian crop.

Let the input pedestrian crop be represented as:

$$X \in Z^{3 \times 192 \times 64}. \quad (1)$$

where 3 denotes the RGB channels, and 192 × 64 denotes the height and width of the resized full-body crop.

The model divides the input image into a fixed 6 × 2 grid. Therefore, the total number of local spatial regions is:

$$N = 6 \times 2 = 12. \quad (2)$$

Each local region is transformed into a token vector with dimension (d = 32). The output of the spatial branch is therefore represented as:

$$S = \{s^1, s^2, \ldots, s^{12}\}, \text{where } s_i \in R^{32}. \quad (3)$$

where s_i denotes the token corresponding to the i-th local body region.

DOI: 10.65322/j.srd.04182026.24

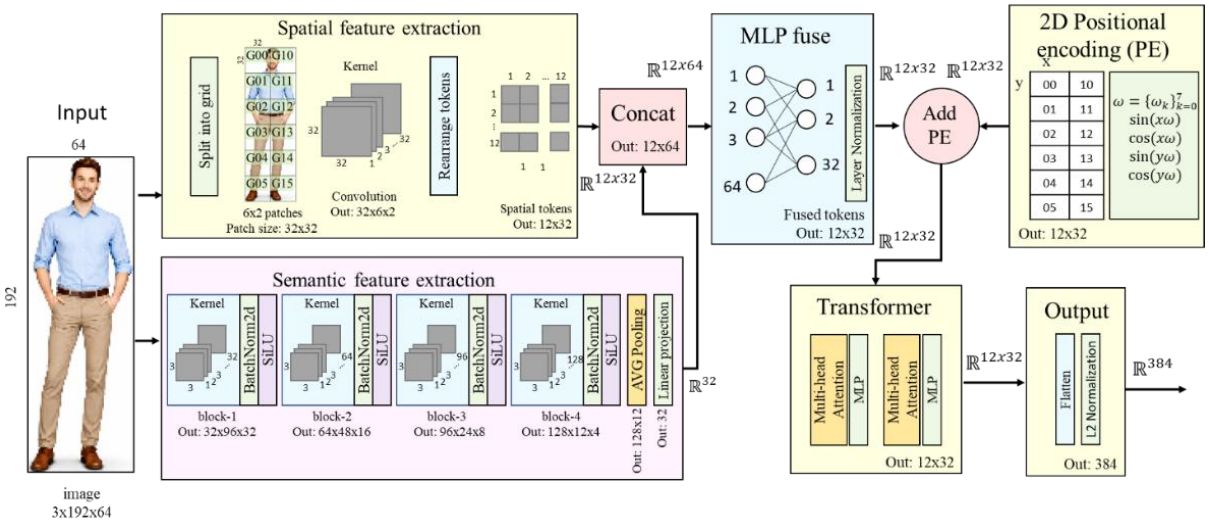


Figure 5. Architecture of the proposed low-resource spatial-semantic transformer feature extractor

In parallel, the semantic branch extracts a global feature vector from the whole pedestrian crop using convolutional blocks, batch normalization, and activation functions. The semantic feature is projected into a 32-dimensional vector:

$$g \in R^{32}. \tag{4}$$

The global semantic vector is concatenated with each local spatial token to form a fused representation:

$$u_i = [s_i \,||\, g]. \tag{5}$$

where || denotes feature concatenation. Since both s_i and g are 32-dimensional, the concatenated vector u_i has 64 dimensions. A multilayer perceptron then projects the fused vector back to the common token dimension:

$$z_i = \mathrm{MLP}(u_i), \text{where } z_i \in R^{32}. \tag{6}$$

where z_i is the fused spatial-semantic token of the i-th body region. The resulting sequence of fused spatial-semantic tokens is:

$$Z = \{z^1, z^2, \ldots, z^{12}\}, \text{where } Z \in R^{12 \times 32}. \tag{7}$$

3.3. Positional Encoding and Transformer Refinement

Because each token corresponds to a fixed body region, positional information must be preserved. Therefore, two-dimensional positional encoding is added to the fused token sequence:

$$\hat{z}_i = z_i + p_i. \tag{8}$$

where p_i is the positional encoding of the i-th spatial region. The position-enhanced token sequence is then processed by lightweight transformer refinement blocks.

For the token matrix (Z), the transformer computes query, key, and value matrices:

DOI: 10.65322/j.srd.04182026.24

$$Q = ZW_Q, K = ZW_K, V = ZW_V. \quad (9)$$

where W_Q, W_K,, and W_V are learnable projection matrices. The self-attention operation is defined as:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \quad (10)$$

where Q is the query matrix, K is the key matrix, V is the value matrix, and d_k is the dimension of the key vector. This operation allows the model to learn relationships among different body-region tokens. For example, upper-body appearance, lower-body clothing, and other local cues can interact during feature refinement.

The transformer output is flattened into a single feature vector:

$$h = \text{Flatten}(Z'). \quad (11)$$

where Z' denotes the refined token sequence. Since the model uses 12 tokens and each token has 32 dimensions, the flattened embedding has:

$$\dim(h) = 12 \times 32 = 384. \quad (12)$$

Finally, L2 normalization is applied:

$$e = \frac{h}{\|h\|_2}. \quad (13)$$

where $e \in R^{384}$ is the final normalized identity embedding, and $\|h\|_2$ is the L2 norm of vector h.

3.4. Similarity Computation

The final identity embedding is used for similarity-based person matching. Since the embedding is L2-normalized, cosine similarity is used to measure the similarity between two pedestrian crops:

$$\text{sim}(e_a, e_b) = e_a^T e_b. \quad (14)$$

where e_a and e_b are normalized embeddings of two pedestrian images. Because the embeddings are L2-normalized, their dot product directly represents cosine similarity. A higher cosine similarity indicates that the two samples are more likely to belong to the same identity.

Cosine distance is defined as:

$$d_{\cos(e_a, e_b)} = 1 - \text{sim}(e_a, e_b). \quad (15)$$

This distance is used to compare anchor-positive and anchor-negative relationships in the embedding space.

3.5. Triplet-Loss-Based Metric Learning

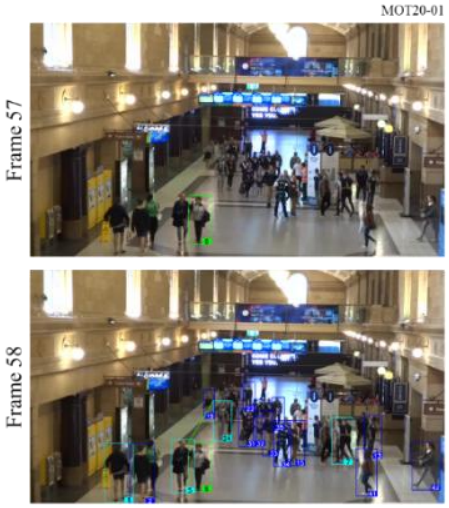
The model is trained using triplet-loss-based metric learning. Each triplet consists of an anchor image x_a, a positive image x_p belonging to the same identity, and a negative image x_n belonging to a different identity. The objective is to reduce the distance between the anchor and positive samples while increasing the distance between the anchor and negative samples.

DOI: 10.65322/j.srd.04182026.24

The cosine triplet loss is defined as:

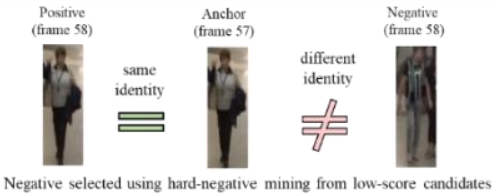
$$L = \max\left(0, d_{\cos(e_a, e_p)} - d_{\cos(e_a, e_n)} + m\right). \quad (16)$$

where e_a, e_p, and e_n are the embeddings of anchor, positive, and negative samples, respectively, and m is the margin parameter. This loss encourages the model to produce embeddings in which same-identity samples are close and different-identity samples are separated.



N	Candidate Track id	Candidate Box (x, y, w, h)	IIS (e10⁻⁹)	ARS (e10⁻⁹)	IIPS (e10⁻⁹)	IISV (e10⁻⁹)	Score (e10⁻⁹)
1	5	655, 783,107,247	93.117	26.623	41.145	489.285	395.670
2	**24**	**849, 605, 84, 185**	**210.811**	**5.767**	**53.906**	**503.662**	**491.700**
3	7	1396, 682, 98, 217	32.258	8.209	342.448	290.772	566.305
4	1	347, 801, 131, 274	182.482	18.281	195.312	369.674	613.582
5	33	1077, 651, 72, 209	71.770	115.324	169.531	306.113	618.803
6	15	1193, 695, 79, 209	71.770	81.831	231.771	277.992	630.491
7	41	1511, 826, 96, 220	18.182	23.458	401.823	134.446	675.328
8	2	463, 802, 114, 278	194.245	49.750	139.323	226.886	700.212
9	34	1128, 696, 78, 214	46.729	95.336	197.656	16.269	732.758
10	13	1550, 672, 80, 200	120	59.821	417.969	289.857	743.877
11	31	967, 640, 80, 176	272.727	5.276	114.323	112.011	779.159
12	42	1770, 795, 129, 223	4.484	118.654	545.313	93.149	849.220
13	19	804, 553, 50, 134	671.642	86.687	21.615	313.803	1112.235
14	35	1120, 588, 58, 152	473.684	78.243	188.281	32.159	1129.988
15	32	1036, 674, 52, 140	600.0	88.393	142.969	20.885	1249.435
16	22	986, 520, 56, 133	684.211	38.769	117.969	6.838	1278.545

Selected negative



Negative selected using hard-negative mining from low-score candidates

Figure 6. Example of triplet generation and hard-negative selection from consecutive MOT20 frames

3.6. Methodological Advantage

The proposed feature extractor is computationally efficient because it uses only 12 spatial-semantic tokens and a compact token dimension of 32. At the same time, it preserves important local body-region information through the 6 × 2 grid and incorporates whole-body semantic appearance through the global branch. Transformer refinement enables interaction among body regions without requiring a large number of image patches. As a result, the model provides a compact 384-dimensional representation suitable for real-time person re-identification in crowded surveillance environments.

Table 5

Proposed model configuration

Configuration item	Value
Input size	3 × 192 × 64
Input image type	Full-body pedestrian crop
Spatial grid size	6 × 2
Number of spatial regions	12
Spatial token dimension	32
Semantic feature dimension	32
Fused token dimension	32
Number of fused tokens	12
Transformer depth	2
Number of attention heads	4

DOI: 10.65322/j.srd.04182026.24

Embedding dimension	384
Embedding normalization	L2 normalization

4. Experimental Setup

This section describes the experimental configuration used to evaluate the proposed low-resource spatial-semantic feature extractor. The evaluation focuses on identity-embedding quality, similarity-based separation between same and different identities, comparison with existing feature extractors, and computational efficiency.

4.1. Dataset Splits

The experiments were conducted using the MOT20-derived full-body pedestrian crop and triplet dataset described in Section 3.1. The dataset is organized into training, validation, query, and gallery subsets. The training and validation subsets were used for metric-learning-based model optimization and validation, while the query and gallery subsets were used for retrieval-style person re-identification evaluation.

The dataset contains 310,935 training images, 36,603 validation images, 465 query images, and 39,423 gallery images. In addition, the triplet annotation files contain 277,688 training triplets, 32,466 validation triplets, and 448 test triplets. The use of hard-negative triplets is important for evaluating identity discrimination in crowded scenes, where different pedestrians may have visually similar appearance.

Table 6

Triplet split configuration used for metric-learning evaluation

Split	Number of triplets	Purpose
Training triplets	277,688	Used for cosine-triplet-loss-based model optimization
Validation triplets	32,466	Used to monitor triplet-learning performance during training
Test triplets	448	Used for final triplet-based identity discrimination evaluation
Total	310,602	Complete triplet annotation set used in the study

The image-level split distribution is presented in Table 4. In addition to the image subsets, the dataset includes triplet annotations designed for metric-learning-based training and evaluation. Each triplet consists of an anchor image, a positive image belonging to the same identity, and a negative image belonging to a different identity. As shown in Table 6, most triplets were assigned to the training subset to support stable optimization of the embedding model. The validation triplets were used to monitor learning behavior, while the test triplets were reserved for final identity-discrimination evaluation.

4.2. Baseline Feature Extractors

To evaluate the effectiveness of the proposed feature extractor, it was compared with widely used convolutional and transformer-based visual feature extractors. The CNN-based baselines include VGG16, ResNet50, ResNet101, and ResNet152. These models are commonly used as deep feature extractors in visual recognition and person re-identification studies. The transformer-based baselines include ViT-B/16, ViT-L/16, and ViT-H/14, which represent different scales of Vision Transformer architectures.

Table 7

DOI: 10.65322/j.srd.04182026.24

Baseline models used for comparison

Model	Model family	Backbone type	Parameters	Model size (MB)	Output feature dimension	Reason for inclusion
VGG16	CNN	Sequential convolutional network	138.358 M	527.792	4096	Classical CNN baseline with high-dimensional fully connected feature representation
ResNet50	CNN	Residual convolutional network	25.557 M	97.492	2048	Widely used CNN backbone for visual feature extraction and ReID comparison
ResNet101	CNN	Deep residual convolutional network	44.549 M	169.942	2048	Deeper residual CNN baseline for evaluating the effect of model depth
ResNet152	CNN	Very deep residual convolutional network	60.193 M	229.617	2048	High-capacity CNN baseline for comparison with compact feature extractors
ViT-B/16	Transformer	Vision Transformer, base model, 16 × 16 patches	86.568 M	330.229	768	Standard transformer baseline for patch-token visual representation
ViT-L/16	Transformer	Vision Transformer, large model, 16 × 16 patches	304.327 M	1160.914	1024	Large transformer baseline for high-capacity visual representation
ViT-H/14	Transformer	Vision Transformer, huge model,	632.046 M	2411.063	1280	Very large transformer baseline for evaluating

DOI: 10.65322/j.srd.04182026.24

		14 × 14 patches				computational -cost difference

All baseline models were evaluated under the same input and embedding-comparison protocol as far as possible. For each model, pedestrian crop images were passed through the feature extraction network, and the resulting feature vectors were used for similarity and distance computation. The proposed method was evaluated as a compact alternative designed specifically for low-resource person re-identification.

4.3. Training Configuration

The proposed model was trained using triplet-loss-based metric learning. Each triplet consists of an anchor image, a positive image from the same identity, and a negative image from a different identity. The model was optimized to reduce the cosine distance between anchor-positive pairs and increase the distance between anchor-negative pairs.

Input pedestrian crops were resized to 192 × 64 pixels. The spatial branch divided each crop into a 6 × 2 grid, producing 12 local spatial tokens. The token dimension was set to 32, and the final embedding dimension was 384. L2 normalization was applied to the final embedding before similarity computation.

The training configuration is summarized in Table 8. The model was trained using AdamW optimization with cosine annealing, mixed-precision computation, and cosine-triplet loss. These parameters are necessary for reproducibility and fair comparison.

Table 8

Training hyperparameters of the proposed model

Hyperparameter	Value
Image preprocessing	Resize to 192 × 64, ImageNet mean-std normalization
Optimizer	AdamW
Learning rate	3×10^{-4}
Weight decay	1×10^{-4}
Batch size	128
Number of epochs	60
Margin value	0.3
Learning-rate scheduler	Cosine annealing
Mixed precision	FP16 automatic mixed precision
Framework	PyTorch and TorchVision
Random seed	42

As shown in Table 8, the model was optimized using AdamW with cosine annealing and cosine-triplet-loss-based metric learning. Mixed-precision training was used to reduce memory consumption and improve training efficiency.

4.4. Evaluation Metrics

The model was evaluated using similarity and distance metrics commonly used in embedding-based person re-identification. Since the proposed model generates L2-normalized embeddings, cosine similarity was used as the main identity-matching score. Cosine distance was also computed to measure separation between same-identity and different-identity samples.

In addition to cosine similarity, several distance and correlation metrics were considered for comparative analysis, including Euclidean distance, Manhattan distance, Minkowski distance, Canberra distance, Pearson correlation, Spearman rank correlation, Kullback–Leibler divergence,

DOI: 10.65322/j.srd.04182026.24

and mean squared error. These metrics provide a broader view of how well the embedding space separates positive and negative pairs.

For threshold-based evaluation, cosine similarity thresholds were varied to identify the best decision boundary between same-identity and different-identity pairs. At each threshold, matching accuracy and classification behavior were analyzed.

4.5. Computational Efficiency Evaluation

Since the proposed model is designed for low-resource person re-identification, computational efficiency was evaluated together with identity-matching performance. The comparison should include parameter count, model size, inference time per image, and frames per second. These indicators are important for real-time surveillance systems, where feature extraction must be performed repeatedly for many pedestrian detections.

The proposed model is expected to require fewer computational resources because it uses only 12 spatial-semantic tokens and a 384-dimensional output embedding. This design reduces token-processing cost compared with standard Vision Transformer models and avoids the heavy feature extraction cost of large CNN backbones.

Table 9

Computational complexity comparison of baseline feature extractors and the proposed model

Model	Inference time (ms/image)	FPS	Computational operations (GMACs)	Computational operations (GFLOPs)
VGG16	9.1302	109.53	3.8820	7.7640
ResNet50	6.3796	156.75	1.0139	2.0278
ResNet101	12.4206	80.51	1.9280	3.8560
ResNet152	18.5923	53.79	2.8433	5.6867
ViT-B/16	19.1041	52.34	11.2863	22.5725
ViT-L/16	40.1511	24.91	39.8560	79.7121
ViT-H/14	117.6359	8.50	108.0731	216.1462
Proposed model	0.7289	1371.97	0.0350	0.0700

Note: The baseline CNN and ViT models are general-purpose visual backbones, whereas the proposed model is a task-specific lightweight person re-identification feature extractor. Therefore, the computational comparison is intended to show deployment efficiency rather than to claim architectural equivalence. FLOPs were calculated from MACs using the approximation 1 MAC ≈ 2 FLOPs.

As shown in Table 9, the proposed model requires only 0.0350 GMACs, corresponding to approximately 0.0700 GFLOPs. This is substantially lower than the evaluated CNN and ViT baselines. The large difference is expected because the proposed model was specifically designed as a compact crop-level person re-identification feature extractor, while the baseline models are larger general-purpose visual backbones. The proposed model reduces computational cost by using a fixed 6 × 2 spatial-token structure, lightweight transformer refinement, and a compact 384-dimensional identity embedding. These results support the suitability of the proposed model for low-resource and real-time person re-identification scenarios.

4.6. Ablation Study Design

An ablation study was conducted to evaluate the contribution of each architectural component. The following variants were evaluated:

1. Spatial branch only, without the semantic branch.
2. Semantic branch only, without local spatial tokens.

DOI: 10.65322/j.srd.04182026.24

3. Spatial-semantic fusion without positional encoding.
4. Spatial-semantic fusion without transformer refinement.
5. Full proposed model.

This analysis shows whether local tokens, global semantic features, positional encoding, and transformer refinement each contribute to the final identity-matching performance.

4.7. Visualization Protocol

To support qualitative evaluation, several visual analyses were included. Example triplets were shown to demonstrate anchor-positive and hard-negative relationships. Pairwise similarity examples were also presented to show how the proposed embedding behaves across consecutive frames. In addition, embedding-space visualization using t-SNE or UMAP can be used to show whether samples of the same identity form compact clusters.

5. Results and Discussion

This section presents the experimental results of the proposed low-resource spatial-semantic feature extractor and discusses its effectiveness for person re-identification in crowded video surveillance. The evaluation focuses on four aspects: similarity-based identity separation, comparison with baseline feature extractors, computational efficiency, and ablation analysis.

5.1. Similarity and Distance-Based Evaluation

The first experiment evaluates how well the proposed embedding separates same-identity and different-identity pedestrian samples. Since the model produces a 384-dimensional L2-normalized embedding, cosine similarity was used as the primary matching score. Higher cosine similarity indicates a higher probability that two pedestrian crops belong to the same identity.

The proposed model showed strong separation between positive pairs and negative pairs. Positive pairs, consisting of pedestrian crops from the same identity, produced higher similarity scores, while negative pairs, consisting of different identities, produced lower similarity scores. This confirms that the proposed spatial-semantic representation can preserve discriminative identity information even when pedestrian samples are extracted from crowded scenes.

In addition to cosine similarity, several distance and correlation metrics were used to analyze the embedding space, including Euclidean distance, Manhattan distance, Minkowski distance, Canberra distance, Pearson correlation, Spearman rank correlation, and mean squared error. These metrics provide complementary evidence about the compactness of same-identity embeddings and the separation of different-identity embeddings.

Table 10

Separation of positive and negative pairs across feature extractors

Metric name	ResNet50	ResNet101	ResNet152	ViT-B/16	ViT-L/16	ViT-H/14	VGG16	Proposed
Euclidean distance	0.447	0.475	0.482	0.603	0.618	0.500	0.407	1.101
Manhattan distance (L1)	7.470	8.888	8.927	12.813	15.325	14.232	11.235	17.305
Minkowski distance (p=3)	0.222	0.226	0.229	0.241	0.230	0.177	0.150	0.475
Canberra distance	262.443	303.674	289.967	191.135	259.332	268.418	397.562	175.616
Cosine similarity	0.299	0.344	0.363	0.459	0.459	0.351	0.243	0.935

DOI: 10.65322/j.srd.04182026.24

Pearson correlation	0.330	0.391	0.412	0.459	0.459	0.351	0.272	0.935
Spearman rank correlation	0.244	0.281	0.274	0.483	0.476	0.361	0.210	0.927
Mean squared error (10^{-4})	2.922	3.358	3.540	11.958	8.971	5.482	1.189	48.685

As shown in Table 10, the proposed model demonstrates the strongest separation for most distance, similarity, and correlation metrics. In particular, the proposed model achieves the highest cosine similarity separation value of 0.935 and the highest Pearson correlation separation value of 0.935, indicating that the learned 384-dimensional embedding provides clear discrimination between same-identity and different-identity pedestrian pairs.

5.2. Cosine-Similarity Threshold Analysis

A threshold-based evaluation was conducted to determine the best decision boundary for identity matching. In this experiment, cosine similarity thresholds were varied, and the matching accuracy was computed at each threshold. If the similarity score between two embeddings exceeded the selected threshold, the pair was classified as the same identity; otherwise, it was classified as a different identity.

The proposed model achieved its best performance at the threshold value of 0.575, where the matching accuracy reached 0.99995. This threshold-based accuracy was computed on the test triplet subset and should be interpreted as pair-level identity discrimination performance rather than full benchmark-level re-identification accuracy.

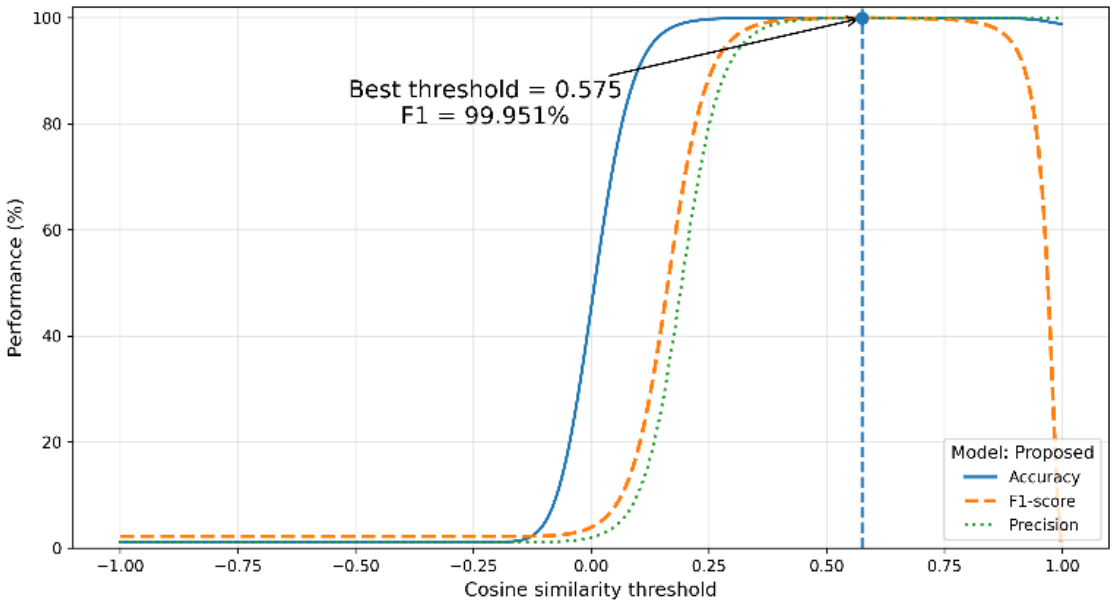


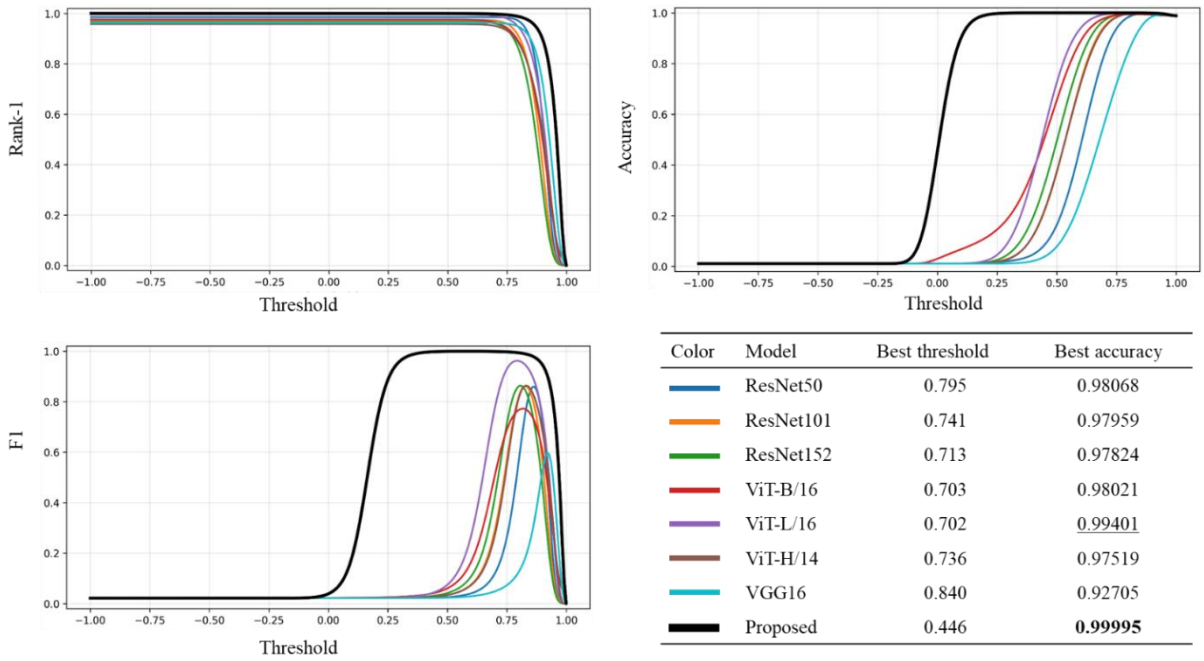
Figure 7. Cosine-similarity threshold analysis for identity matching

5.3. Comparison with Baseline Feature Extractors

DOI: 10.65322/j.srd.04182026.24

The proposed method was compared with widely used CNN-based and transformer-based feature extractors, including VGG16, ResNet50, ResNet101, ResNet152, ViT-B/16, ViT-L/16, and ViT-H/14. These models were selected because they represent different levels of depth, parameter count, feature representation capacity, and computational cost.

CNN-based models such as VGG16 and ResNet variants provide strong visual representations, but their performance depends heavily on backbone depth and feature dimensionality. ResNet50 provides a practical balance between accuracy and computational complexity, while ResNet101 and ResNet152 increase representational capacity at the cost of higher computational demand. VGG16 is useful as a classical CNN baseline but is less efficient because of its large parameter count.



Color	Model	Best threshold	Best accuracy
	ResNet50	0.795	0.98068
	ResNet101	0.741	0.97959
	ResNet152	0.713	0.97824
	ViT-B/16	0.703	0.98021
	ViT-L/16	0.702	0.99401
	ViT-H/14	0.736	0.97519
	VGG16	0.840	0.92705
	Proposed	0.446	**0.99995**

Figure 8. Performance comparison of baseline feature extractors and the proposed model

Transformer-based baselines, including ViT-B/16, ViT-L/16, and ViT-H/14, provide strong global token interaction. However, they require a larger number of image patches and significantly higher computational resources. This makes them less suitable for real-time crowded surveillance settings, where the feature extractor must be repeatedly applied to many detected pedestrian crops.

Compared with these baselines, the proposed model uses only 12 spatial-semantic tokens and produces a compact 384-dimensional embedding. This design reduces computational cost while preserving local body-region structure and global appearance information. The results show that the proposed method provides a favorable balance between identity-discrimination capability and computational efficiency.

5.4. Computational Efficiency Analysis

Computational efficiency is a critical requirement for person re-identification in real-time surveillance systems. In crowded scenes, a large number of pedestrian crops may be generated from each frame, and the feature extractor must process them quickly. Therefore, model size, parameter count, inference time, and frames per second were analyzed.

DOI: 10.65322/j.srd.04182026.24

The proposed model is computationally efficient because it avoids large CNN backbones and reduces transformer token processing to only 12 spatial-semantic tokens. In contrast, large CNNs and standard ViT architectures require more parameters and higher memory usage. ViT-L/16 and ViT-H/14 are especially computationally expensive and may be impractical for deployment in low-resource systems.

The proposed model achieved an inference time of 0.73 ms per image and a processing speed of 1372 FPS, with a model size of 899kB and 235680 parameters. These results demonstrate that the proposed model is suitable for resource-constrained surveillance applications where both speed and identity-matching accuracy are important.

5.5. Ablation Study

An ablation study was conducted to evaluate the contribution of each component of the proposed architecture. The tested variants include: spatial branch only, semantic branch only, spatial-semantic fusion without positional encoding, spatial-semantic fusion without transformer refinement, and the full proposed model.

To evaluate the contribution of each architectural component, an ablation study was conducted by progressively modifying the spatial branch, semantic branch, positional encoding, and transformer refinement module.

Table 11

Ablation study results of the proposed model components

Model variant	Spatial branch	Semantic branch	Fusion	Positional encoding	Transformer refinement	Rank-1 (%)	mAP (%)	Score
A0	Yes	No	No	Yes	No	92.59	47.38	0.8499
A1	Yes	Yes	No	Yes	No	93.52	52.99	0.8663
A2	Yes	Yes	Yes	No	Yes	93.03	58.50	0.8788
A3	Yes	Yes	Yes	Yes	No	93.15	58.20	0.8784
A4	Yes	Yes	Yes	Yes	Yes	93.02	58.71	0.8793

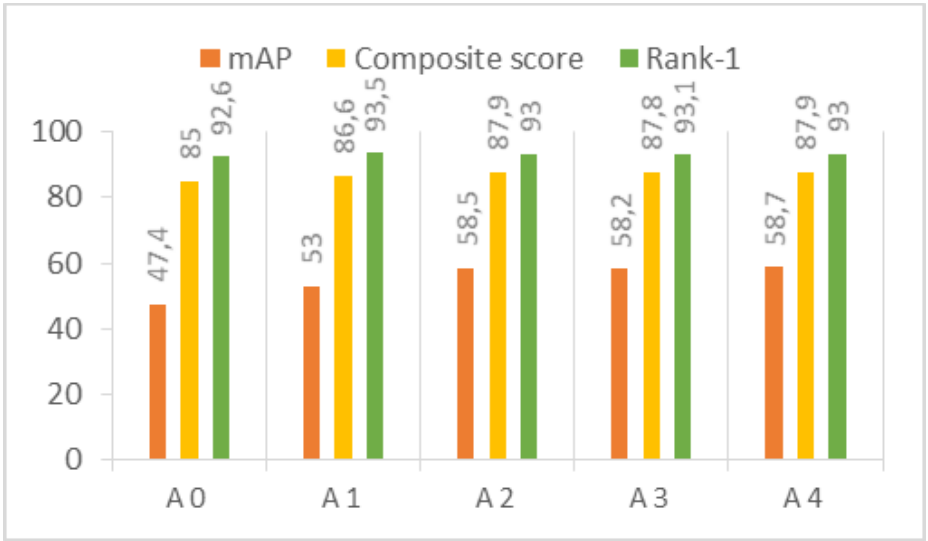


Figure 9. Ablation study of the proposed spatial-semantic feature extractor

As shown in Table 11 and Figure 9, adding semantic information improves performance compared with the spatial-only variant. The full proposed model achieves the highest composite score and the highest mAP, confirming the benefit of combining spatial tokens, semantic features, positional encoding, and transformer refinement in a unified compact representation.

DOI: 10.65322/j.srd.04182026.24

The spatial-only variant evaluates the contribution of local body-region tokens. The semantic-only variant evaluates whether global appearance information alone is sufficient for identity matching. The variant without positional encoding tests whether the fixed body-region order contributes to performance. The variant without transformer refinement evaluates the importance of token interaction among body regions.

The full model achieved the best overall performance, confirming that all components contribute to the final identity representation. The spatial branch preserves local body information, the semantic branch provides whole-crop appearance context, positional encoding preserves region order, and transformer refinement models relationships among body regions.

5.6. Qualitative Analysis

Qualitative analysis was performed using example triplets and pairwise similarity visualization. In anchor-positive-negative examples, the anchor and positive samples belong to the same identity, while the negative sample belongs to a different identity but may have visually similar appearance. The proposed model assigns higher similarity to anchor-positive pairs and lower similarity to anchor-negative pairs, demonstrating its ability to distinguish visually similar pedestrians.

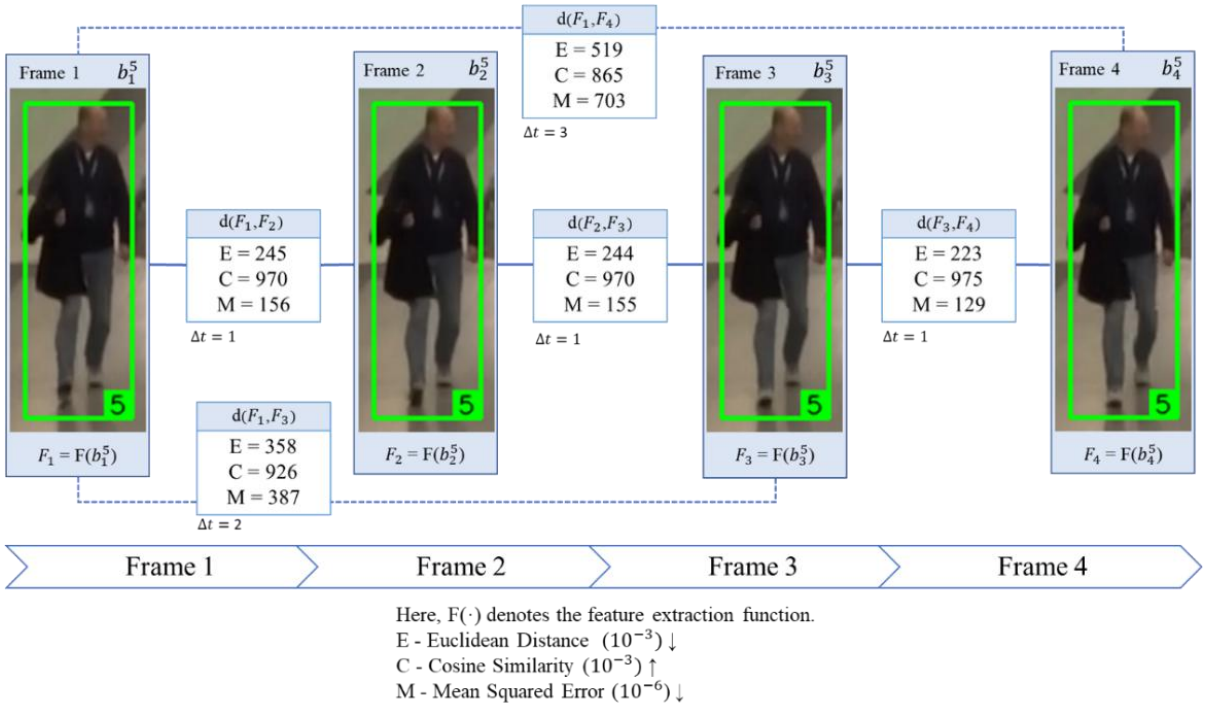


Figure 10. Pairwise similarity examples for same-identity and different-identity pedestrian crops

Pairwise similarity examples across consecutive frames also show that the proposed embedding remains stable for the same identity despite small changes in pose, scale, and visibility. This stability is important for crowded video surveillance because identity matching must remain reliable across sequential frames and partial occlusions.

To qualitatively examine the discriminative structure of the learned embedding space, t-SNE was applied to selected 384-dimensional embeddings generated by the proposed feature extractor.

DOI: 10.65322/j.srd.04182026.24

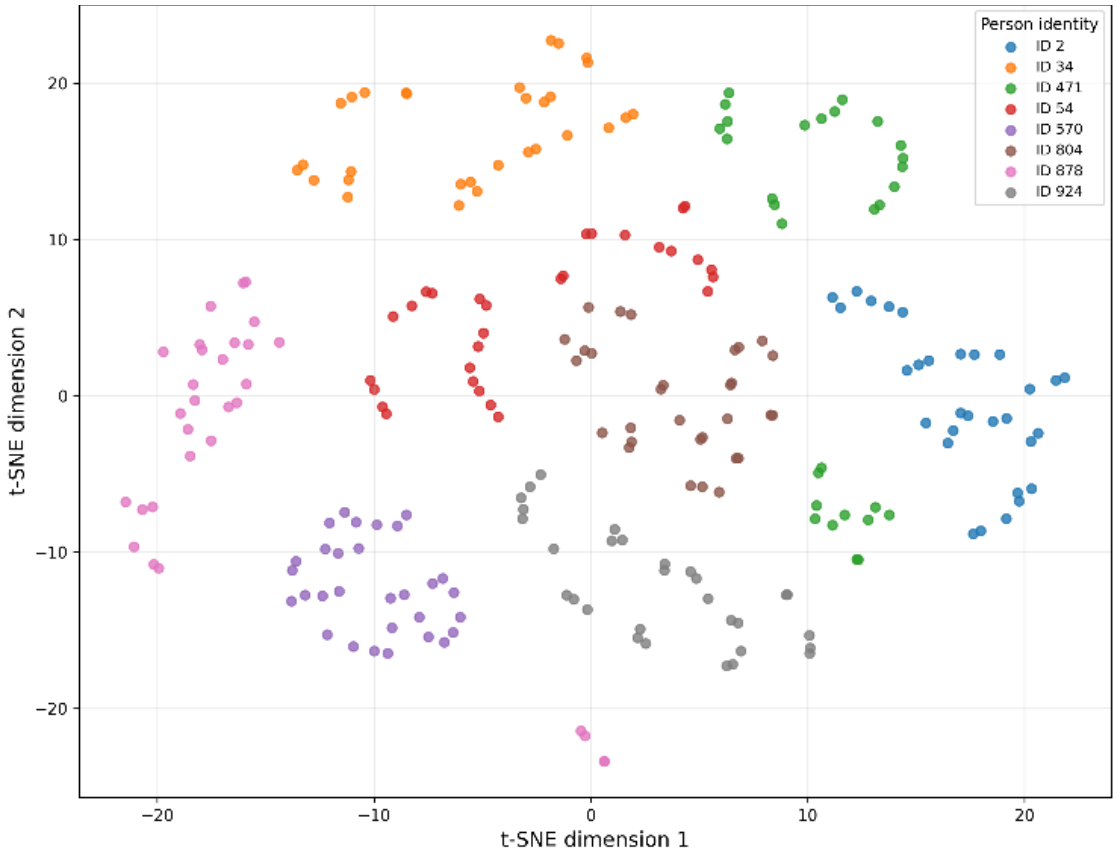


Figure 11. Embedding-space visualization of identity clusters using t-SNE

As shown in Figure 11, samples belonging to the same identity generally form local groups in the projected embedding space. Some identities appear as several sub-clusters, which can be explained by pose variation, scale changes, partial occlusion, and frame-level appearance differences in crowded MOT20 scenes. Overall, the visualization confirms that the proposed embedding preserves identity-related structure.

5.7. Discussion

The experimental results demonstrate that the proposed low-resource spatial-semantic feature extractor is effective for person re-identification in crowded surveillance scenes. The main advantage of the proposed method is its ability to combine local body-region information with global semantic appearance while maintaining low computational complexity. Unlike standard vision transformers that process many image patches, the proposed model uses only 12 spatial-semantic tokens. This makes the transformer refinement stage more efficient.

The results also indicate that local and global information are complementary. Local tokens help preserve body-region details that may remain visible under partial occlusion, while the semantic branch provides whole-body context. The ablation study supports this observation because the full model outperforms variants that use only one branch or remove transformer refinement.

Compared with large CNN and ViT baselines, the proposed model provides a more practical balance between performance and efficiency. This is important for real-time applications, where the feature extractor must process multiple pedestrian detections per frame. Therefore, the proposed method is suitable for surveillance systems that require identity matching under resource constraints.

However, the method also has limitations. The dataset is derived from MOT20 sequences and therefore inherits the crowded-scene characteristics and annotation constraints of the source

DOI: 10.65322/j.srd.04182026.24

benchmark. In addition, the fixed 6 × 2 grid assumes a consistent full-body crop structure. If crops are severely misaligned or contain incomplete body regions, some local tokens may become less informative. Future work may improve robustness by adding adaptive region weighting, occlusion-aware token selection, or temporal feature aggregation across video frames.

6. Conclusion

This paper proposed a low-resource spatial-semantic transformer feature extractor for person re-identification in crowded video surveillance. The proposed model was designed to generate compact and discriminative identity embeddings while reducing computational complexity. Each full-body pedestrian crop is divided into a fixed 6 × 2 spatial grid, producing 12 local body-region tokens. These local spatial tokens are fused with a global semantic feature extracted from the whole crop, enhanced with positional encoding, and refined using lightweight transformer blocks.

The final output of the model is a 384-dimensional L2-normalized embedding that can be directly used for cosine-similarity-based identity matching and triplet-loss-based metric learning. This design allows the model to preserve local body-region information, incorporate global pedestrian appearance, and model relationships between body regions without requiring a large number of image patches or a heavy backbone architecture.

The proposed method was evaluated using a MOT20-derived full-body pedestrian crop and triplet dataset prepared for person re-identification. The dataset contains 387,426 curated pedestrian crops, 2,089 identities, and 310,602 triplet annotations, including hard-negative samples for metric-learning evaluation. The experimental analysis showed that the proposed spatial-semantic representation is suitable for separating same-identity and different-identity samples while maintaining a compact embedding structure and low computational requirements.

Compared with conventional CNN and transformer-based feature extractors, the proposed model offers a practical balance between identity-discrimination capability and efficiency. This makes it suitable for real-time or resource-constrained surveillance systems where multiple pedestrian detections must be processed repeatedly. The results also confirm the importance of combining local body-region cues with global semantic appearance for robust person re-identification in crowded scenes.

References

1. Z. Xue *et al.*, “FDD-YOLO: A Lightweight Multi-scale Prohibited Items Detection Model,” *EAI Endorsed Trans. AI Robot.*, vol. 4, pp. 1–15, 2025.
2. K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.
3. K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc.*, pp. 1–14, 2015.
4. G. Xu, A. Sohail Khan, A. J. Moshayedi, X. Zhang, and Y. Shuxin, “The Object Detection, Perspective and Obstacles In Robotic: A Review,” *EAI Endorsed Trans. AI Robot.*, vol. 1, 2022.
5. A. Dosovitskiy *et al.*, “An Image Is Worth 16X16 Words: Transformers for Image Recognition At Scale,” *ICLR 2021 - 9th Int. Conf. Learn. Represent.*, 2021.
6. J. Zhou, S. Zhao, S. Li, B. Cheng, and J. Chen, “Research on Person Re-Identification through Local and Global Attention Mechanisms and Combination Poolings,” *Sensors*, vol. 24, no. 17, p. 5638, 2024.
7. A. Asperti, L. Naldi, and S. Fiorilla, “An Investigation of the Domain Gap in CLIP-Based Person Re-Identification,” *Sensors*, vol. 25, no. 2, p. 363, 2025.

DOI: 10.65322/j.srd.04182026.24

8. R. Lin, R. Wang, W. Zhang, A. Wu, Y. Sun, and Y. Bi, "Person Re-Identification Method Based on Dual Descriptor Feature Enhancement," *Entropy*, vol. 25, no. 8, p. 1154, 2023.

9. S. Li, L. Sun, and Q. Li, "CLIP-ReID: Exploiting Vision-Language Model for Image Re-Identification without Concrete Text Labels," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023, vol. 37, no. 1, pp. 1405–1413.

10. K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang, "Omni-Scale Feature Learning for Person Re-Identification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 3702–3712.

11. K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang, "Learning Generalisable Omni-Scale Representations for Person Re-Identification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 5056–5069, 2022.

12. H. Luo *et al.*, "A Strong Baseline and Batch Normalization Neck for Deep Person Re-identification," *IEEE Trans. Multimed.*, vol. 22, no. 10, pp. 2597–2609, 2020.

13. 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 480–496.

14. S. He, H. Luo, P. Wang, F. Wang, H. Li, and W. Jiang, "TransReID: Transformer-Based Object Re-Identification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 15013–15022.

15. M. Ye, S. Chen, C. Li, W.-S. Zheng, D. Crandall, and B. Du, "Transformer for Object Re-Identification: A Survey." 2024.

16. B. Hu, X. Wang, and W. Liu, "PersonViT: Large-scale Self-supervised Vision Transformer for Person Re-Identification." 2024.

17. G. Liu, K. Xu, J. Zhu, Y. Ge, and X. Chen, "A Local-Global Transformer-Based Model for Person Re-Identification," *PLoS One*, vol. 20, no. 11, p. e0335848, 2025.

18. X. Zhang *et al.*, "TE-TransReID: Towards Efficient Person Re-Identification via Local Feature Embedding and Lightweight Transformer," *Sensors*, vol. 25, no. 17, p. 5461, 2025.

19. G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou, "Learning Discriminative Features with Multiple Granularities for Person Re-Identification," in *Proceedings of the 26th ACM International Conference on Multimedia*, 2018, pp. 274–282.

20. W. Qin, Y. Li, J. Zhang, X. Wen, J. Guo, and Q. Guo, "Attention-Based Hybrid Contrastive Learning for Unsupervised Person Re-Identification," *Sci. Rep.*, vol. 15, p. 13238, 2025.

21. C. Zhao, Y. Qin, B. Zhang, Y. Zhao, and B. Wu, "An End-to-End Occluded Person Re-Identification Network with Smoothing Corrupted Feature Prediction," *Artif. Intell. Rev.*, vol. 58, no. 2, p. 53, 2025.

22. H. M. Zangana, F. M. Mustafa, and M. Omar, "A Hybrid Approach for Robust Object Detection: Integrating Template Matching and Faster R-CNN," *EAI Endorsed Trans. AI Robot.*, vol. 3, pp. 1–10, 2024.

23. R. S. M. Saad, M. M. Moussa, N. S. Abdel-Kader, H. Farouk, and S. Mashaly, "Deep Video-Based Person Re-Identification (Deep Vid-ReID): Comprehensive Survey," *EURASIP J. Adv. Signal Process.*, vol. 2024, no. 1, p. 63, 2024.

24. Z. Zhang, D. He, S. Liu, B. Xiao, and T. S. Durrani, "Completed Part Transformer for Person Re-Identification," *IEEE Trans. Multimed.*, vol. 26, pp. 2303–2313, 2024.

25. H. Ni, Y. Li, L. Gao, H. T. Shen, and J. Song, "Part-Aware Transformer for Generalizable Person Re-Identification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 11246–11255.

DOI: 10.65322/j.srd.04182026.24

26. C. Yu, X. Liu, Y. Wang, P. Zhang, and H. Lu, "TF-CLIP: Learning Text-Free CLIP for Video-Based Person Re-Identification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, vol. 38, no. 7, pp. 6764–6772.

27. C. Zhu, W. Zhou, and J. Ma, "Person Re-Identification Network Based on Edge-Enhanced Feature Extraction and Inter-Part Relationship Modeling," *Appl. Sci.*, vol. 14, no. 18, p. 8244, 2024.

28. P. Dendorfer *et al.*, "MOT20: A benchmark for multi object tracking in crowded scenes," *arXiv*, pp. 1–7, Mar. 2020.