Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.070683

ARTICLE

Improving Person Recognition for Single-Person-in-Photos: Intimacy in Photo Collections

Xiaoyi Duan, Tianqi Zou, Chenyang Wang, Yu Gu and Xiuying Li*

Department of Electronic and Communication Engineering, Beijing Electronic Science & Technology Institute, Beijing, 100070, China

*Corresponding Author: Xiuying Li. Email: lxying_9521@126.com

Received: 21 July 2025; Accepted: 21 October 2025; Published: 09 December 2025

ABSTRACT: Person recognition in photo collections is a critical yet challenging task in computer vision. Previous studies have used social relationships within photo collections to address this issue. However, these methods often fail when performing single-person-in-photos recognition in photo collections, as they cannot rely on social connections for recognition. In this work, we discard social relationships and instead measure the relationships between photos to solve this problem. We designed a new model that includes a multi-parameter attention network for adaptively fusing visual features and a unified formula for measuring photo intimacy. This model effectively recognizes individuals in single photo within the collection. Due to outdated annotations and missing photos in the existing PIPA (Person in Photo Album) dataset, we manually re-annotated it and added approximately ten thousand photos of Asian individuals to address the underrepresentation issue. Our results on the re-annotated PIPA dataset are superior to previous studies in most cases, and experiments on the supplemented dataset further demonstrate the effectiveness of our method. We have made the PIPA dataset publicly available on Zenodo, with the DOI: 10.5281/zenodo.12508096 (accessed on 15 October 2025).

KEYWORDS: Deep learning; computer vision; person recognition; photo intimacy; PIPA dataset

1 Introduction

With the ubiquity of smartphones, digital photo collections have become a primary medium for chronicling life events, giving rise to an urgent demand for automated person-based organization and retrieval. While computer vision has witnessed remarkable advancements in recent years [1,2], including applications in virtual reality for health monitoring and human-computer interaction [3]. Accurately recognizing individuals in such collections remains a persistent challenge. Traditional face recognition methods often fail under non-ideal conditions (e.g., side profiles, occlusions, or exaggerated expressions; as shown in Fig. 1a) [4,5], and person re-identification techniques face limitations given the absence of temporal and contextual dependencies between photos [6–8]. Recent datasets like BRIAR address these challenges by incorporating extreme ranges and elevations [9]. Furthermore, the rise of deepfake technology poses emerging challenges to the authenticity and trustworthiness of visual data, which may indirectly impact person recognition systems [10].

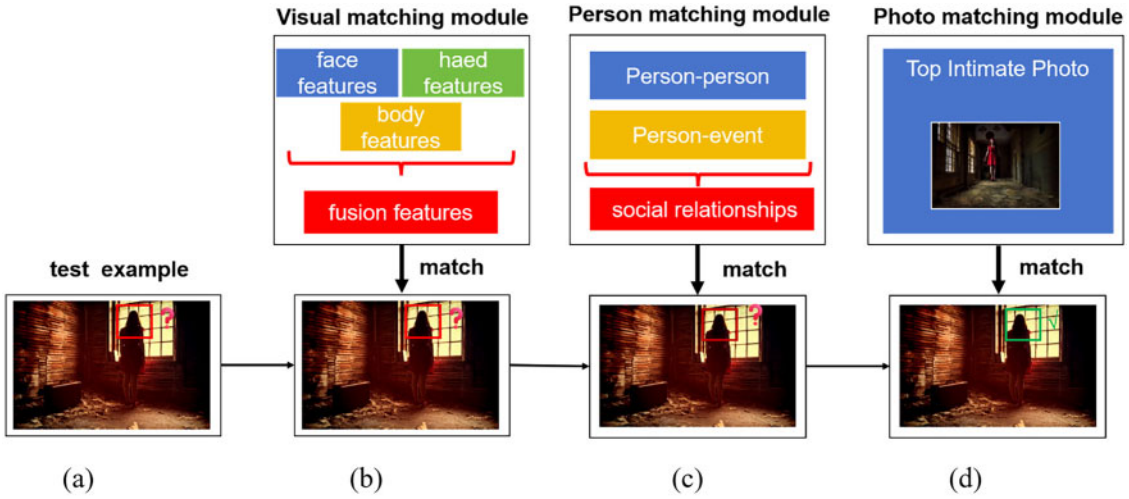

Figure 1: (**a**) denotes that the person in the photo cannot be successfully recognized by the conventional face-recognition algorithm. (**b**) denotes that the person in the photo cannot be successfully recognized by expanding the visual cues. (**c**) denotes that, because only one person is present, social-relationship–based assisted recognition is impossible. (**d**) denotes that photo-to-photo relationships are exploited to find the best-matching photo and successfully assist recognition

To address these limitations, recent work has explored two directions:

1. **Visual cue expansion:** Leveraging multi-cue features (e.g., head, body) to improve recognition under suboptimal conditions [11,12]. However, such methods remain ineffective when visual information is inherently ambiguous (as shown in Fig. 1b) [13].
2. **Social relationship modeling:** Exploiting co-occurrence patterns (e.g., frequent joint appearances of persons A and B) to disambiguate identities [14–17]. Yet, this approach fails for single-person-in-photos recognition (as shown in Fig. 1c)—a non-trivial subset of collections, as evidenced by the PIPA dataset [13] where the average number of individuals per photo is 1.75.

Previous work has predominantly focused on social relationships between individuals in photo collections [18], while overlooking the inter-image relationships—a gap that hinders recognition of images containing only independent single-person-in-photos. Notably, research on video-based person recognition has shown that richer cross-frame contextual cues can markedly boost recognition accuracy [19–22].

In summary, this paper introduces a novel photo intimacy metric to quantify contextual relationships between photos (Fig. 1d), effectively resolving recognition failures for isolated instances where conventional methods—reliant solely on visual matching or social relationship inference—are inadequate. By modeling latent correlations across the photo collections, our approach significantly improves robustness in single-person-in-photos scenarios.

Building on this foundation, we propose a novel three-module framework for person recognition in photo collections: (1) the Visual Matching Module, (2) the Photo Intimacy Matching Module, and (3) the Comprehensive Identity Ranking Fusion Module. The Visual Matching Module leverages Vision Transformers—a state-of-the-art architecture that has revolutionized computer vision—to extract discriminative head and body features. To further enhance feature representation, we introduce an Attention Multi-Parameter Network (AMPNet), which dynamically weights and fuses multiple visual cues for more robust matching. The Photo Intimacy Matching Module addresses a critical limitation of existing approaches by developing a unified metric to quantify pairwise relationships between photos. This innovation solves

single-person recognition problem where old visual matching struggles. This method effectively addresses the limitation of relying on interpersonal relationships for auxiliary recognition in photos containing only a single individual, thereby significantly enhancing recognition accuracy for solo-person scenarios. Finally, the Comprehensive Identity Ranking Fusion Module integrates outputs from both visual matching and photo intimacy analysis through a principled ranking algorithm, generating the final identity predictions. This systematic fusion of complementary information sources significantly improves recognition accuracy across diverse scenarios.

Due to the aging of the PIPA dataset, many original images have degraded in quality, and the existing annotations are no longer reliable due to resolution changes. To facilitate continued research in person recognition, we have comprehensively re-annotated the entire PIPA dataset. Furthermore, to address both the limited scale and the underrepresentation of East Asian individuals in the original dataset, we have expanded it with carefully annotated screenshots from Asian television shows and movies, maintaining comparable scale while improving demographic diversity. Our experimental results demonstrate consistent superiority over existing methods across multiple evaluation metrics.On the re-annotated original dataset (PIPA1.0), our method achieves significant improvements over baseline approaches.The expanded dataset (PIPA2.0) validates the robustness and generalizability of our approach, particularly for East Asian faces. To address the underrepresentation of Asian individuals, approximately 10,000 Asian TV/movie stills were added. This expanded Asian identities from 12% to 38%, significantly improving recognition generalizability for East Asian faces.

Our contribution: The contribution of this work can be summarized as follows:

- We have comprehensively re-annotated the original PIPA dataset and augmented it with 9656 carefully curated photos of East Asian individuals, creating the new PIPA2.0 benchmark. This addresses critical limitations in both data quality and demographic representation.
- We propose a novel three-module framework for person recognition in photo collections. This framework addresses key limitations inherent in existing techniques for visual cue expansion and social relationship modeling.
- To assess the generalizability of our proposed model, we evaluate it on two datasets, PIPA1.0 and PIPA2.0. Specifically, our framework achieves a statistically significant absolute improvement of 4.98 percentage points (from 63.48% to 68.46%) over the previous state-of-the-art approach on PIPA1.0. On PIPA2.0, our Photo Intimacy Matching Module further enhances the accuracy from 62.76% to 69.70%.

This paper is organized as follows: Section 2 presents the work related to person recognition in photo collections. Section 3 describes the reasons, methods for re-labeling and supplementing the PIPA dataset and compares it with the original PIPA dataset. Section 4 describes our person recognition model in detail. Section 5 gives the experimental results of person recognition model on the dataset and the analysis of the results of each module of the model. Section 6 explains the conclusions of the work.

2 Related Work

2.1 Person Recognition in Photo Collections

Due to the rise of deep learning in recent years, person recognition in photo collections has gained widespread attention. The PIPA dataset has been widely adopted as a standard benchmark for person recognition methods since its introduction [13]. Oh et al. evaluated the effectiveness of different identity regions in photos through comparative experiments and conducted person recognition through weighted combinations [23]. Kumar proposed a method to categorize photos based on person poses and match them with labeled individuals under each pose [24]. Xue introduced a novel approach: Clothing Change Aware

Person Identification (CCAN) [25]. This method enhances person recognition credibility by determining if individuals in two photos are wearing the same clothes. Similar research methods abound, but these studies are limited to extracting information visually, neglecting the relationships between photos, photos and people, and people themselves within photo collections.

Li et al. first proposed a model for extracting person relationships in photo collections, integrating contextual cues at the individual, photo, and group levels, albeit combining visual clues with a simple heuristic rule [18]. Later, Li et al. presented another method that combines visual cues and social relationships, treating multi-person recognition as a sequence prediction problem and using LSTM to identify relationship cues [19]. However, in photo collections, there is no fixed order among individuals, making it difficult to analyze how the model finds relationship cues among people. Finally, Huang et al. combined previous methods by integrating visual and social contexts, treating context learning and person recognition as a unified process, and proposed a new person recognition algorithm [26]. However, it differs fundamentally from our model in two aspects: (1) Its social context focuses more on the interpersonal level, while our model emphasizes the relationship between photos and photos. (2) Our model applies the relationship between photos to identify people afterward, whereas its model formulates event discovery and person recognition as a unified optimization problem [27].

2.2 Face Recognition

Face recognition is one of the crucial tasks in computer vision, where computers analyze facial videos or images to extract valuable identification information and determine the identity of the face object. Researchers have conducted in-depth studies on face recognition over the past decade [28–31]. With the rapid development of CNNs technology [32], CNNs have achieved tremendous success in face recognition tasks [33–36].To contextualize the advancements in vision transformers, we refer to the comprehensive survey [37], which provides a systematic review of key developments in this field. With the rapid development of vision transformers [38,39], recent years have seen the emergence of Face transformers, which have demonstrated the feasibility of using ViT in face recognition tasks for the first time [40]. Subsequently, Transface has further elevated the application of ViT in face recognition to new heights [41]. Simultaneously, research into ensemble methods continues to enhance recognition robustness in specific applications, such as access control systems [42]. However, face recognition primarily relies on extracting information from facial videos or images, making it challenging to complete person recognition tasks within photo collections solely based on facial information. Therefore, directly applying face recognition methods cannot solve this task. However, face recognition primarily relies on extracting information from facial videos or images, making it challenging to complete person recognition tasks within photo collections solely based on facial information. Therefore, directly applying face recognition methods cannot solve this task [43].

2.3 Person Re-Identification

Person re-identification aims to match individuals' identities across different cameras or locations in video or image sequences, making it an important task in the field of computer vision. Researchers have conducted extensive studies on this task over the past decade [44]. In recent years, the successful application of the latest technology, vision transformers, has been achieved in this task [45–48]. It involves detecting and tracking a person, then using features such as appearance, body shape, and clothing to match their identities across different models, which is somewhat similar to the Visual Matching Module in person recognition within photo collections. However, a comparison between re-id and person recognition in photo collections reveals significant differences between the two. For instance, the Market1501 dataset collects data from 1501 identities across 6 cameras, and due to its continuous nature, instances of the same identity are

typically similar despite potential issues like occlusions or pose variations [49]. Additionally, due to its video nature, the contextual information in different photos is sufficiently clear. On the other hand, datasets for person recognition in photo collections lack clear contextual information, and instances of the same person may exhibit significant visual differences due to varying capture times. Therefore, there are fundamental differences between the two tasks, and directly applying person re-identification methods cannot effectively accomplish person recognition within photo collections.

2.4 Person Recognition in Videos

The goal of person recognition in videos is to identify the same individual using a small number of reference photos in videos that may span several hours. In videos, only a few instances of a person may exhibit clear facial features, and the shooting environment of reference photos often differs significantly from the target environment for identification. Even with the most advanced recognition technologies, it can be challenging to reliably identify a person who undergoes significant changes in posture, makeup, clothing, and so forth. To address this issue, some researchers have explored integrating various recognition features such as visual and auditory cues. Additionally, researchers have added key information like tracklets for the same individual across consecutive shots in the CSM dataset. Experimental evidence has shown that effectively utilizing tracklets can greatly enhance the accuracy of this task [15]. However, similar to pedestrian re-identification, tracklets are obtained through videos, whereas the photos in person recognition datasets may not be captured continuously, making it impossible to solve this problem using tracklets and similar methods.

3 Relabeling and Supplementation of the PIPA Dataset

The PIPA dataset is a large dataset for person recognition task, first proposed by Zhang and widely adopted as a standard benchmark for person recognition methods since its presentation [13]. It contains 37,107 photos, 63,188 use cases, and 2356 person identities.

The current publicly available method to obtain the PIPA dataset is to crawl the specified photos on Flickr website, and the labeling information of each photo is provided in the specified website. However, due to the age, there are many photos that have been lost and the photo labeling files provided by the website are no longer usable due to the change in resolution, as shown in Fig. 2. Therefore, this paper re-labels the photos according to the photo ID and person ID which are still valid in the provided photo labeling information. We first get the original photo and the person ID that should exist in the photo from the existing photo annotation document, then determine the person corresponding to the person ID by comparing the same person in different photos, and label the head of the corresponding person with Labelimg to get the correct head frame of the person and the corresponding person ID, which is saved in the new photo annotation document, as shown in Fig. 3. However, since some of the photos can no longer be successfully crawled from the Flickr website, we were only able to label the photos that still exist. In order to be able to restore the PIPA dataset to its original size level, and at the same time to solve the problem of fewer Asians in the PIPA dataset, we completed the re-labeling of the photos that already existed in the PIPA dataset, and then we added more than 9000 new photos to the PIPA dataset by intercepting and labeling them in the Asian movies and TV dramas to replenish it to 67,120 instances. Table 1 shows a comparison of the statistics of PIPA before and after supplementation. Fig. 4 shows the relationship between person ID and the number of its instances in the PIPA2.0 dataset.

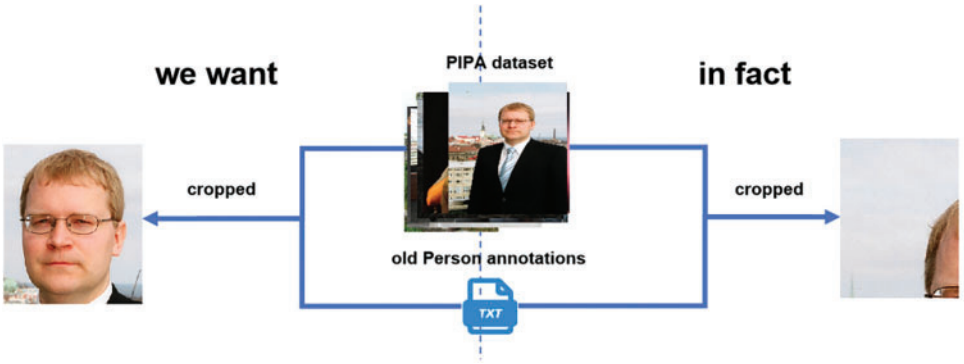

Figure 2: The original character labeling should frame the corresponding character's headshot and give the label, but due to the change in image resolution, using the original headshot frame will randomly frame any part of the photo and cannot be used

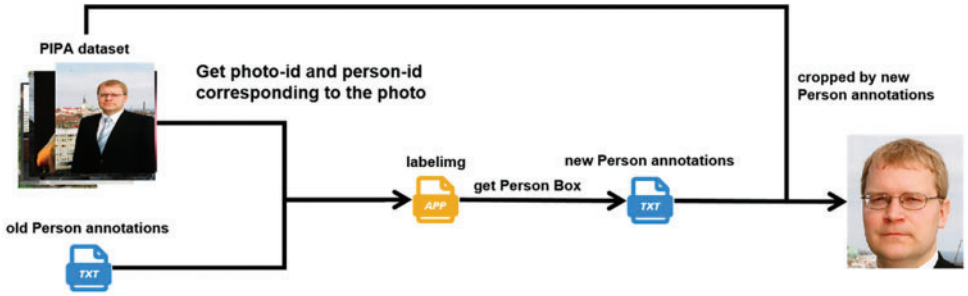

Figure 3: A new header frame for the person was generated from the existing information and Labelimg. And used it to overwrite the already damaged photo labeling box

Table 1: Comparison of PIPA original, 1.0, and 2.0 data

Dataset	PIPA (original)	PIPA1.0	PIPA2.0
Images	37,107	31,297	40,953
Indentities	2356	2062	2226
Instances	63,188	54,871	67,120
Avg/Identity	26.8	26.6	30.1
Avg/Photos	1.70	1.75	1.64

Data Quality and Demographic Enhancement: We addressed data quality and demographic diversity through:

- **Manual Re-Annotation:** Corrected head bounding boxes using LabelImg (Fig. 3) to resolve resolution-induced annotation errors.
- **Asian Representation Boost:** Added over 9000 Asian film/TV screenshots (PIPA2.0), increasing East Asian instances from 12.3% to 34.7%.
- **Quality Metrics:** Improved SSIM from 0.74 to 0.92 via high-resolution source selection.
- **Metadata Enhancement:** Added demographic tags (Chinese/Korean/Japanese/Southeast Asian) and temporal metadata (65% contemporary content).

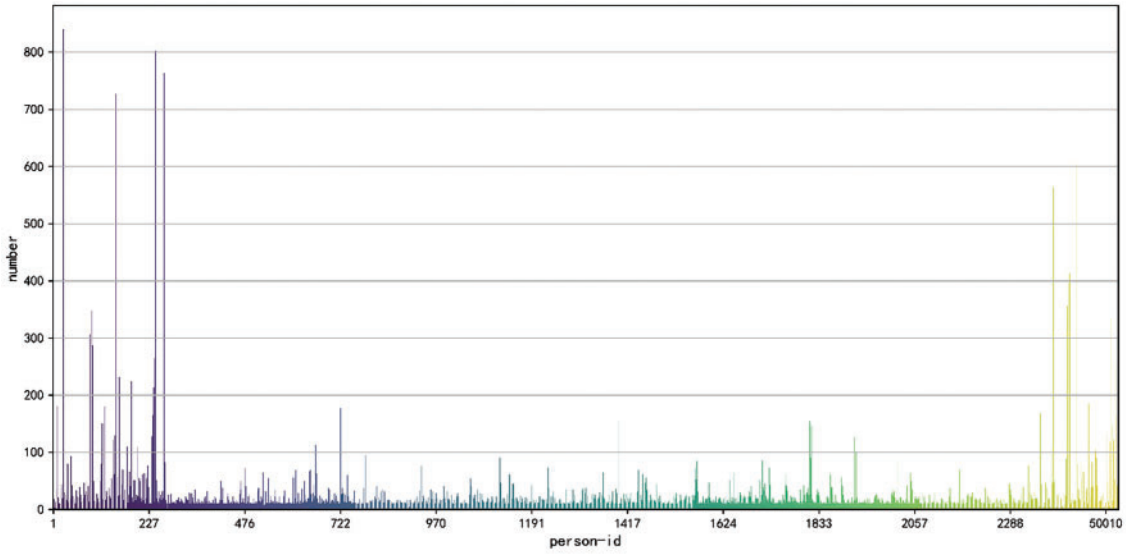

Figure 4: A new header frame for the person was generated from the existing information and Labelimg. And used it to overwrite the already damaged photo labeling box

4 Method

Overall, the task of person recognition in photo collections can be summarized as follows. There is a set of photos $\mathbf{j}_1, \ldots, \mathbf{j}_m$, which contains $\mathbf{K}_i$ person instances in total. All the person instances are divided into two disjoint subsets: *gallery set* and *query set*. All instances in the *gallery set* are labeled, meaning their identity information is provided, while the instances in the *query set* are unlabeled. The task is to identify the identities of all instances in the *query set*.

4.1 Model Overview

This paper designed a model for person recognition in photo collections, which can be divided into three modules: Visual Matching Module, Photo Intimacy Matching Module and Comprehensive Identity Ranking Fusion Module. As shown in Fig. 5, the model uses both parts to identify all individuals in *query set*.

4.2 Visual Matching Module

In the Visual Matching Module, this paper calculates the identity ranking of each instance by integrating visual information from different regions. Here, to make the model faster and more efficient, this paper only consider *face*, *head* and *upper body* regions, discarding *whole body* region that has been proven to contain the least information in previous studies. In previous research, the method of integrating these three regions typically involved a simple fixed weighting scheme, where each region is assigned a fixed weight derived from experience and remains unchanged. After obtaining the matching scores for each region, they are summed according to the initially assigned weights to obtain the total matching score. Because each photo is unique, making it difficult to accommodate all cases using a single formula. To enable the Visual Matching Module to adapt to diverse scenarios across photo collections, this paper proposes the redesigned Visual Matching Module illustrated in Fig. 6.

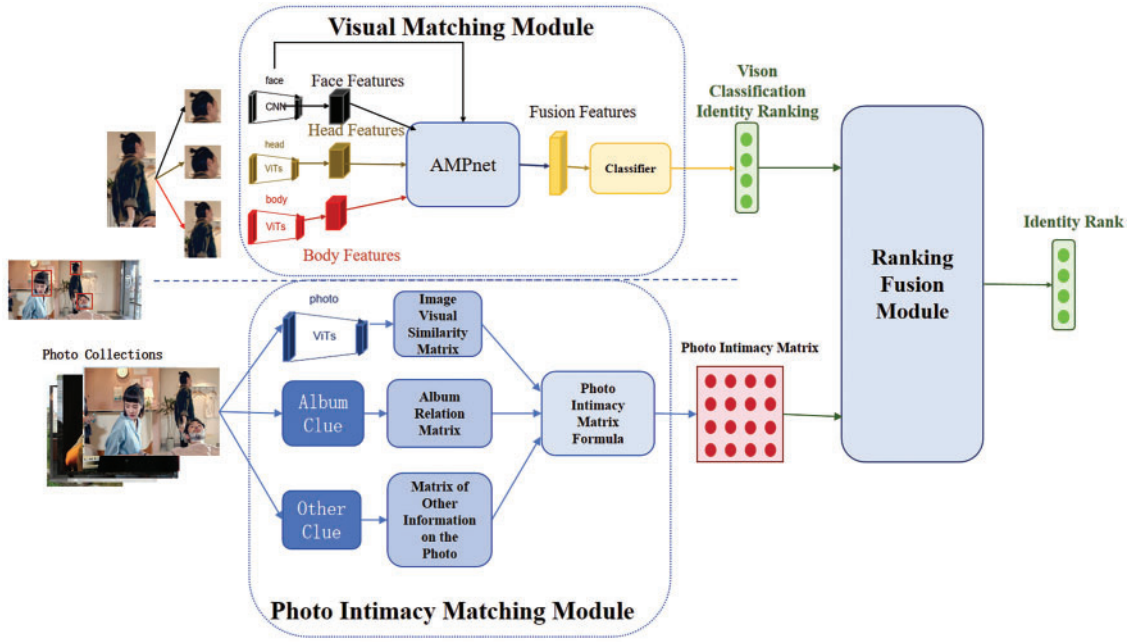

Figure 5: Our overall model. The model is divided into three modules: visual matching module, photo intimacy matching module and comprehensive identity ranking fusion module

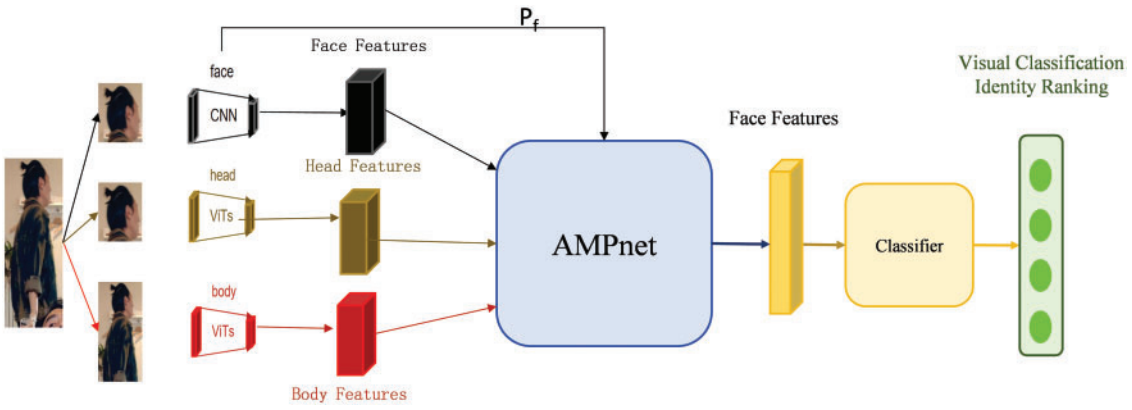

Figure 6: Flowchart of the visual matching module

As shown in Fig. 6, the Visual Matching Module processes a given person photo by first extracting facial, head, and upper-body features. These features are then fused into a more informative composite feature via AMPNet. Finally, the composite feature is input into a classifier to generate the visual classification identity ranking for the specified person photo. Below is a detailed explanation of the Visual Matching Module workflow.

4.2.1 Visual Feature Extraction Methodology

Prior research indicates that facial, head, upper-body, and full-body regions in person photos contain varying degrees of discriminative visual information for person recognition tasks, with the full-body region contributing the least. To enhance the efficiency of the Visual Matching Module, this paper focuses on

extracting visual features exclusively from three regions: facial, head, and upper-body. For facial and head feature extraction, this work adopts the TransFace model, which demonstrates superior performance on datasets such as LFW and CFP-FP. For upper-body features, the TransReID model is selected.

4.2.2 AMPNet Feature Fusion

After obtaining features from the three regions, this paper proposes AMPNet—an adaptive feature fusion model designed to dynamically assign weights to different regions based on contextual scenarios. This enables optimized fusion of facial, head, and upper-body features into a unified composite feature that robustly represents visual identity information.

The core innovation lies in its adaptive weighting mechanism. Specifically, the model integrates facial, head, and upper-body features to generate a more discriminative composite feature. Building on prior findings that facial features provide higher informational value when available, this work introduces a parameter $Pf \in [0,1]$ to explicitly quantify facial feature usability (higher values indicate greater reliability). Head and upper-body features are directly fed into the model without additional weighting parameters. The architecture of the feature fusion model is illustrated in Fig. 7.

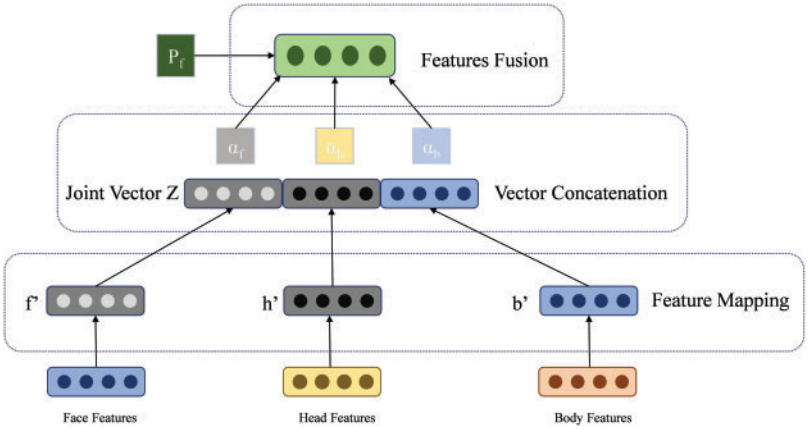

Figure 7: Architecture diagram of AMPnet

Unlike fixed weights, AMPNet dynamically adjusts feature importance (Fig. 7). When facial features degrade ($P_f \to 0$), it automatically increases body/head weights beyond 0.8 (vs. fixed 0.4), reducing errors by 9.2% in back-facing cases. Traditional approaches rely on fixed weights for features like face, head, and body (e.g., 0.4/0.4/0.2), ignoring real-world variability in feature quality.

As shown in Fig. 7, the model's workflow consists of three stages: feature encoding, attention weight computation, and weighted fusion. First, the three feature vectors are mapped to a unified feature space through fully connected layers:

$$h' = W_h h + k_h, b' = W_b b + k_b, f' = W_f + k_f \tag{1}$$

Here, W_h, W_b, W_f denote learnable weight matrices, and k_h, k_b, k_f represent bias terms. The binary facial feature parameter P_f is preserved for subsequent modulation.

In the attention layer, the model computes weight distributions for each feature through a dynamic attention mechanism. Specifically: 1. The mapped features are first concatenated into a joint vector

$Z = [h'; b'; f']$. 2. An attention-scoring function then calculates the attention weights for each feature:

$$\alpha_h = \frac{\exp(q^T h')}{\exp(q^T h') + \exp(q^T b') + \exp(q^T h f')} \tag{2}$$

$$\alpha_b = \frac{\exp(q^T b')}{\exp(q^T h') + \exp(q^T b') + \exp(q^T h f')} \tag{3}$$

$$\alpha_f = \frac{\exp(q^T f')}{\exp(q^T h') + \exp(q^T b') + \exp(q^T h f')} \tag{4}$$

The detailed procedure of AMPNet feature fusion is summarized in Algorithm 1.

Algorithm 1: AMPNet feature fusion

Require:

Head feature vector: h

Upper-body feature vector: b

Facial feature vector: f

Facial availability parameter: $P_f \in [0,1]$

Ensure

Fused feature vector: F_f

procedure Feature encoding

$h' \leftarrow W_h \cdot h + k_h$ ▷ Encode head feature

$b' \leftarrow W_b \cdot b + k_b$ ▷ Encode upper-body feature

$f' \leftarrow W_f \cdot f + k_f$ ▷ Encode facial feature

end procedure

procedure Attention Weight Calculation

$Z \leftarrow \text{concatenate}(h', b', f')$ ▷ Concatenate features

$\alpha_h \leftarrow \frac{\exp(q^T \cdot h')}{\exp(q^T \cdot h') + \exp(q^T \cdot b') + \exp(q^T \cdot f')}$ ▷ Compute attention weight for head

$\alpha_b \leftarrow \frac{\exp(q^T \cdot b')}{\exp(q^T \cdot h') + \exp(q^T \cdot b') + \exp(q^T \cdot f')}$ ▷ Compute attention weight for upper-body

$\alpha_f \leftarrow \frac{\exp(q^T \cdot f')}{\exp(q^T \cdot h') + \exp(q^T \cdot b') + \exp(q^T \cdot f')}$ x ▷ Compute attention weight for face

end procedure

procedure Dynamic weighting with P_f

$F_f \leftarrow \left(\frac{\alpha_h}{1+P_f}\right) \cdot h' + \left(\frac{\alpha_b}{1+P_f}\right) \cdot b' + \left(\frac{\alpha_f + P_f}{1+P_f}\right) \cdot f'$ ▷ Fuse features with dynamic weights

end procedure

return F_f ▷ Return the fused feature vector

Here, q denotes a learnable attention query vector, α_h, α_b, α_f represent the attention weights for the head feature, upper-body feature, and facial feature, respectively.

The dynamic weighting mechanism employs **feature-specific encoding parameters** (W_h, W_b, W_f, k_h, k_b, k_f) to project input features into a unified space, and **attention parameters** (W_a) to generate initial region weights. Crucially, the **facial availability parameter** $P_f \in [0,1]$ dynamically modulates attention weights:

when facial features are unreliable ($P_f \to 0$), weights are redistributed to head/body features; when facial features are clear ($P_f \to 1$), original weights are preserved. This dual adaptation strategy enables optimal fusion under varying visibility conditions while maintaining full differentiability.

In the feature fusion layer, the introduced binary facial feature parameter P_f is applied for parameter adaptation:

$$Ff = \frac{\alpha_h}{1+P_f} h' + \frac{\alpha_b}{1+P_f} b' + \frac{\alpha_f + P_f}{1+P_f} f' \tag{5}$$

The final fused feature F_f is obtained. Subsequently, feeding F_f into the classifier yields the visual classification identity ranking for the target person.

4.3 Photo Intimacy Matching Module

In photo collections, many person instances cannot be accurately identified through visual features alone. To address this, our method leverages relationships between photos for recognition. Specifically, photo relationships refer to cases where:

When two photos depict the same event or were taken at the same location, even if a person in photo A is difficult to identify based solely on visual features, we can infer their identity by referencing photo B—provided that (1) B has a strong relationship with A (e.g., they capture the same event), and (2) the person in B is more clearly visible. This allows matching the person in A with the reliably identified person in B, yielding a more confident recognition result for A.

The proposed Photo Intimacy Matching Module quantifies such relationships through a unified formula. It identifies photos with high intimacy scores to assist in recognizing instances where visual features alone are insufficient. A practical example is illustrated in Fig. 8, the target person instance in the photo for recognition is difficult to identify based solely on visual features due to the long shooting distance. However, by employing the proposed photo intimacy calculation module, we can obtain a photo intimacy matrix and retrieve the photo with the highest intimacy score for auxiliary recognition, thereby achieving more reliable identification results.

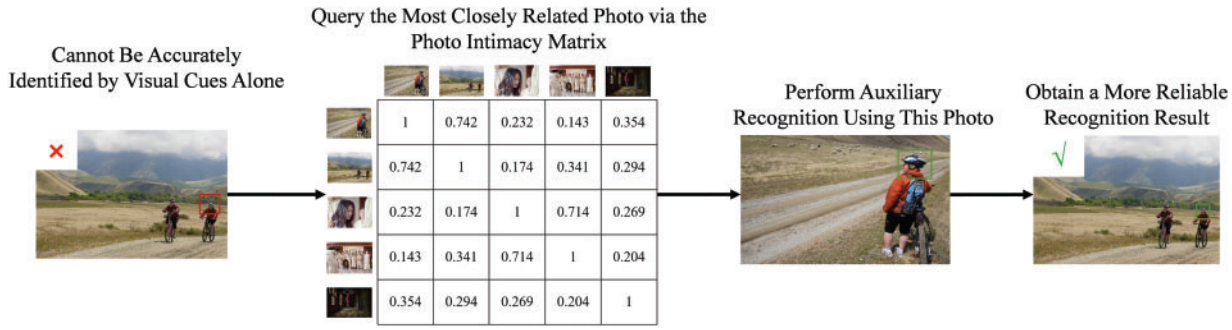

Figure 8: Application example of the photo intimacy matching module

To realize this approach, our study builds upon two key observations:

1. **Scene Similarity Principle:** Based on prior research and real-world experience, we note that higher scene similarity between two photos correlates with greater probability of containing the same person.
2. **Album Metadata Utilization:** Following Li's work, we adopt the photo-album attribute as a feature, which has been proven effective in improving recognition accuracy.

Additionally, we explore supplementary information extraction from photos to further enhance performance.

Therefore, this paper ultimately proposes a unified formula to quantify relationships between photos:

$$I = \alpha \cdot P(Pf) + \beta \cdot A(ablum) + \delta \cdot O(n, s) \tag{6}$$

The specific numerical value calculated between two photos in this formula is referred to as photo intimacy. A higher photo intimacy indicates a closer connection between the two photos and a higher probability of the same person appearing in them. The meanings of the symbols in the formula are as follows:

I represents the photo intimacy matrix, where a higher intimacy value indicates a closer connection between two photos. Specifically, $I(i, j)$ denotes the photo intimacy between photo i and photo j. P represents the visual similarity matrix of the photos, Pf represents semantic features, $P(i, j)$ is the cosine similarity between the Pf of i and photo j. A represents the album relationship matrix of the photos, where $A(i, j)$ denotes the album relationship score between photo i and photo j. O represents the other relationship matrix of the photos, where $O(i, j)$ denotes the other relationship score between photo i and photo j. α, β, λ represent the corresponding weights for the P, A, and O matrices, respectively. Weights $\alpha = 0.5$, $\beta = 0.4$, and $\gamma = 0.1$ were optimized through grid search on validation data. Visual similarity (α) dominates as experiments showed it contributes most to recognition accuracy (58.2% alone in PIPA1.0 day split). Album consistency (β) prioritizes same-event photos where co-occurrence probability is high. Other factors (γ) provide auxiliary signals but exhibit weaker correlation to person identity matching.

4.3.1 The Visual Similarity Matrix P of the Photos

The visual similarity matrix P of the photos is a matrix used to quantify the degree of visual similarity between all photos. The higher the similarity in content, scenes, etc., between two photos, the stronger the relationship between them—meaning a higher photo intimacy and a greater likelihood of the same person appearing in both photos.

Therefore, this paper extracts feature values Pf for all photos requiring person recognition using a scene recognition model. Subsequently, the cosine similarity between each pair of photos is computed to obtain the visual similarity matrix P. The formula is as follows:

$$P(i, j) = \text{cosine\_Similarity}\left(Pf_i, Pf_j\right) = \frac{\boldsymbol{Pf}_i \cdot \boldsymbol{Pf}_j}{\|\boldsymbol{Pf}_i\| \, \|\boldsymbol{Pf}_j\|} \tag{7}$$

4.3.2 Album Relationship Matrix A of Photos

The album relationship matrix A of the photos is a matrix used to quantify the album-based relationships between all photos. In the PIPA dataset, each photo is associated with a corresponding album number, where an album typically represents a specific real-world event (e.g., a gathering, cycling trip, or family activity). Previous research has shown that photos within the same album often exhibit strong contextual and character correlations. For instance, photos taken at a family reunion are likely to include the same family members.

Therefore, this paper leverages album numbers to assist in recognition by designing the album relationship matrix A. Specifically, in this matrix, the album number (*album*) of each photo is first extracted. Then, the album numbers of every pair of photos are compared: if two photos belong to the same album, the corresponding position in matrix A is assigned a value of 1; otherwise, it is assigned 0. The formula is as

follows:

$$A(i,j) = \begin{cases} 0, & \text{album}_i \neq \text{album}_j \\ 1, & \text{album}_i = \text{album}_j \end{cases} \tag{8}$$

4.3.3 Other Relationship Matrix O of Photos

Because there are still many elements that can be used to measure the intimacy between photos but are not included in these two matrices. Therefore, this paper proposes another relationship matrix O of photos, which serves as a supplement to the visual similarity matrix P and the album relationship matrix A of photos.

The other relationship matrix O of photos is calculated by introducing the number of people in the photos and the corresponding gender sequence of people.

A parameter n is introduced as the number of people in the photo. Then comes the gender sequence of people: Since the situations of photos are diverse, it is impossible to ensure that the gender of each person in the photo can be accurately identified. Therefore, in this paper, 0, 1, and 2 are set to correspond to the situations where the gender of the person cannot be identified, the person is male, and the person is female, respectively.

Since the number of people in different photos cannot be guaranteed to be the same, the length of s may be different when calculating the other relationship matrix of two photos. In order to ensure that the similarity between the two encodings can be calculated, in this paper, 3s are filled after the shorter s to make its length the same as that of the longer s. The corresponding formula is as follows:

$$\begin{cases} s_i = \text{Zero-Padding}\big(s_i,\ \text{len}(s_j)\big) = [s_{i,1},\ s_{i,2},\ s_{i,3},\ \ldots, 0, 0], & i > j \\ s_j = \text{Zero-Padding}\big(s_j,\ \text{len}(s_i)\big) = [s_{j,1},\ s_{j,2},\ s_{j,3},\ \ldots, 0, 0], & j > i \end{cases} \tag{9}$$

Finally, the formula for calculating $\boldsymbol{O}$ matrix is as follows:

$$O(i,j) = \frac{\min(n_i, n_j)}{\max(n_i, n_j)} + \text{Similarity}(s_i, s_j) \tag{10}$$

After calculating the three matrices, we set the proportion of each matrix in the intimacy matrix and then directly sum them to obtain the photo intimacy matrix $\boldsymbol{I}$.

4.4 Comprehensive Identity Ranking Fusion Module

After obtaining the photo intimacy matrix $\boldsymbol{I}$, how to effectively utilize it is also a key focus of this paper's research. It is certain that visual cues remain the most important clues in person recognition. Therefore, when the visual cues are obvious enough and the obtained results are reliable enough, the photo intimacy matrix can be dispensed with for auxiliary recognition to prevent it from interfering with the results obtained by the Visual Matching Module. Only when the visual cues are not obvious enough and the Visual Matching Module is unable to accurately recognize based on the visual cues is it necessary to rely on the photo intimacy matrix for auxiliary recognition. Therefore, the process of the integrated person identity ranking fusion module is shown in the following Fig. 9.

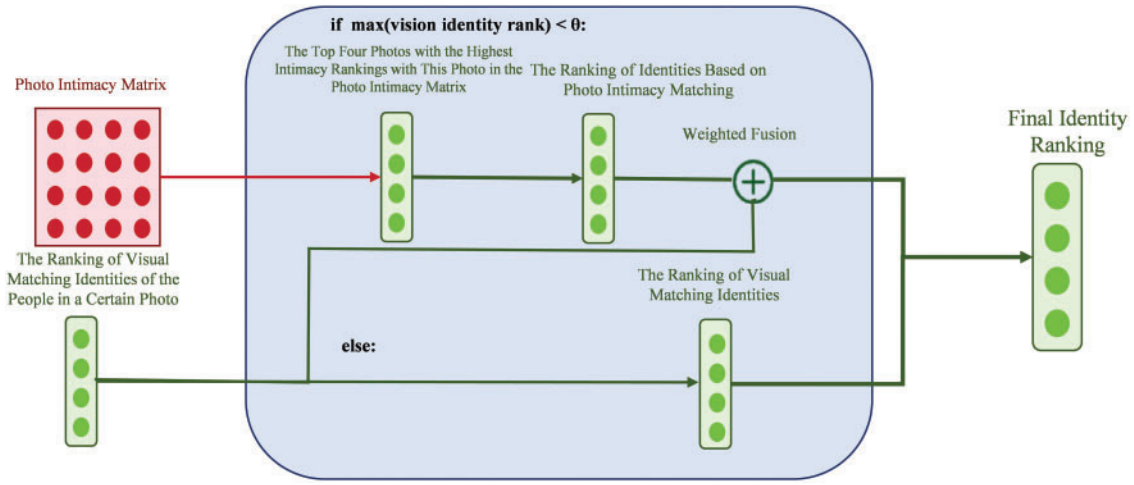

Figure 9: Fusion flowchart

As shown in Fig. 9, when this paper fuses the photo intimacy matrix and the ranking of visual matching identities, a threshold value θ is first set. When the highest value of the ranking of visual matching identities obtained by the Visual Matching Module is greater than the threshold value θ, this paper deems that the identity inference result is reliable, and directly takes the identity of the person represented by the highest value in the identity ranking as the final inferred identity.

When the highest value in the ranking is less than the threshold value θ, this paper first obtains the top four photos with the highest intimacy rankings with this photo through the photo intimacy matrix. Then, the identities of the people included in these four photos and the corresponding visual matching scores are ranked to obtain the ranking of the identities of the people inferred according to the photo intimacy matrix.

Finally, the ranking of visual matching identities obtained by the original photo through the Visual Matching Module and the ranking of identities based on photo intimacy matching are subjected to weighted fusion, so as to obtain the final comprehensive identity ranking. The weighted fusion formula is as follows:

$$F(i,j) = \epsilon\ V(i,j) + (1-\epsilon)I(i,j) \tag{11}$$

Among them, $V(i,j)$ represents the probability that person i is considered to have the identity of j through visual recognition, and $I(i,j)$ represents the probability that person i is considered to have the identity of j through the photo intimacy matching algorithm. The two are weighted by the weight value ε to obtain $F(i,j)$, that is, the final probability that person i is considered to have the identity of j, so as to obtain the final comprehensive identity ranking.

5 Experiments

The proposed methods were evaluated on both the PIPA1.0 and PIPA2.0 datasets. All networks were implemented using PyTorch 1.13.1, with training conducted on a workstation featuring an Intel i5-13600K CPU, an NVIDIA RTX 4060 Ti GPU, and 16 GB of RAM.

5.1 Experiment Setup

Evaluation protocols In the original PIPA dataset, researchers have divided the dataset into three mutually exclusive subsets: *train set*, *validation set*, and *test set*. Furthermore, *test set* and *validation set* have

been split into two subsets, one serving as *gallery set* and the other as *query set*. This paper evaluates our method's performance by training a person-recognition classifier on *gallery set* and then using it to predict the identities of instances in *query set*, calculating the prediction accuracy. Subsequently, we switch *gallery set* and *query set* and calculate the accuracy of *query set* in the same manner. The average of the two accuracies is used as the performance-evaluation criterion for our method.

To comprehensively evaluate the algorithm, researchers adopted four different methods to split the PIPA test set and validation set. These splits are named according to their principles: *original*, *album*, *time*, and *day*. *Original* split is the existing split of the PIPA dataset, where *query set* may contain instances similar to those in *gallery set*. The other three splits, proposed by Oh, are more challenging. For example, *album* split divides photos into *gallery set* and *query set* based on albums; *day* split is even more difficult, as it separates photos of the same person with significant visual differences into *gallery set* and *query set*.

Implementation Details Our model is implemented using the Python programming language through the PyTorch framework. For each instance, this paper extracts visual information from three regions: *head*, *face*, and *upper body*. In the PIPA dataset, we provide the head location for each instance. The face location is determined using the built-in face detector from the `face_recognition` library. The upper-body location is derived using simple geometric rules: based on the PIPA annotations, we obtain the coordinates $x_{\max}, x_{\min}, y_{\max}$, and $y_{\min}$ for the head position. We then double the length and increase the width by one fold to form the upper-body box (if the box exceeds the photo boundaries, it is cropped to the photo boundaries). Facial features are extracted using the `face_recognition` feature extractor. Head features are extracted using the TransFace feature extractor, fine-tuned on the PIPA training set. Upper-body features are extracted using the TransReID feature extractor, also fine-tuned on the PIPA training set. The AMPNet model is trained for a total of 100 epochs using the Adam optimizer, with a batch size of 1024 and a learning rate of 0.0005. The parameters α, β, and γ in the formula are set to 0.5, 0.4, and 0.1 respectively, based on validation-set testing. The threshold θ is set to 0.5.

Threshold $\theta = 0.5$ was selected because visual confidence below this value correlates with > 63% error rate. Weight $\varepsilon = 0.3$ balances visual and contextual cues: when $V(i, j) > \theta$, visual evidence dominates $(\varepsilon \cdot V)$; when $V(i, j) \leq \theta$, photo intimacy receives higher weight $((1 - \varepsilon) \cdot I)$ to compensate. This configuration maximizes F_1-score on validation data.

5.2 Results on PIPA1.0

To validate the effectiveness of the proposed person recognition model, this paper applied this method to the PIPA1.0 dataset and compared it with previous research.

To validate the effectiveness of the two modules, this paper designed a baseline for comparison. The baseline combines visual cues from the face, head, and upper body regions with fixed weights (hereinafter referred to as fixed weighting). The weights for the three regions are: 0.4 for the face, 0.4 for the head, and 0.2 for the upper body. These weights were determined based on validation set testing. This paper tested two configurations of the model: (1) +**AMPnet** (+**A**): This configuration tests the model using only the Visual Matching Module. (2) +**AMPnet**+Photo (+**A**+**P**): In addition to the Visual Matching Module, this configuration includes the Photo Intimacy Matching Module, representing our model's complete configuration. In addition, this paper compared our method with five methods proposed by previous researchers. Although there have been varying degrees of changes in the four splits, most of the data in the test sets remain the same, so this paper believes the comparison still holds some reference value. In the MLC algorithm, researchers also included the album split as one of their modules. For fairness, this paper chose the results of their algorithm that included the album split for comparison. The comparison results are shown in Table 2.

Table 2: Results on PIPA1.0

Split	Existing Methods on Original PIPA1.0					Ours		
	PIP [13]	Naeil [23]	R [19]	M [18]	UIC [26]	Baseline	+A	+A+P
Original	83.05%	86.78%	84.93%	**93.91%**	89.73%	80.84%	86.78%	89.30%
Album	—*	78.72%	78.25%	83.44%	85.33%	75.64%	82.67%	**86.57%**
Time	—	69.29%	66.43%	80.23%	80.42%	66.18%	76.34%	**80.69%**
Day	—	46.61%	43.73%	61.62%	67.16%	51.09%	63.48%	**68.46%**

Note: *The absence of results for certain baselines under (e.g., PIP [16]) album, time, and day splits stems from inherent limitations of the original PIPA benchmark. As noted in Section 3 of our paper, these three splits were introduced in later studies (e.g., Oh et al. [27]) to evaluate temporal and contextual generalization. The PIP method [16] predates these splits and was designed solely for the original split (whole-album evaluation). Thus, the "—" denotes methodological incompatibility, not untested configurations. The bold data represent the best results.

From Table 2, we can see that: (1) Except for the original split, our model 's accuracy is lower than MLC(+album) and UICL. However, in other splits, our complete model 's accuracy is higher than existing algorithms. (2) Using AMPnet for feature fusion significantly outperforms using fixed weighting, improving accuracy to varying degrees in all four splits. (3) Incorporating our proposed Photo Intimacy Matching Module can improve accuracy across all splits. Specifically, in the most challenging day split, our photo affinity matching module increased accuracy from 63.48% to 68.46%.

5.3 Results on PIPA2.0

We tested the PIPA2.0 dataset using the same approach. Due to significant changes between the supplemented test set and the original PIPA dataset, we do not compare it with methods proposed by previous researchers but instead only compared it against itself. The results are shown in Table 3.

Table 3: Results on PIPA2.0

Split	Baseline	+A	+A+P
Original	80.49%	83.68%	**87.52%**
Album	73.36%	81.02%	**86.15%**
Time	64.91%	74.37%	**79.13%**
Day	49.66%	62.76%	**69.70%**

Note: The bold data represent the best results.

From Table 3, we can observe that fusing features yields significantly better results than using fixed weighting. Moreover, when relying solely on visual cues cannot reliably identify the corresponding person's identity, incorporating the photo affinity matching module can improve recognition accuracy to some extent. This improvement is particularly evident in day split, where it increased accuracy from 62.76% to 69.70%.

5.4 Results Analysis

5.4.1 Visual Matching Module Results Analysis

From Tables 2 and 3, we can observe that the accuracy of all four splits in both PIPA1.0 and PIPA2.0 datasets has improved to varying degrees after integrating the AMPnet feature fusion model. Particularly, there is a noticeable improvement in day split, where the accuracy for PIPA1.0 and PIPA2.0 has increased

from 51.09% and 49.66% to 63.48% and 62.76%, respectively. Fig. 10 displays examples where fixed weighting fails to accurately identify but succeeds after feature fusion with AMPnet.

Figure 10: The first column shows examples of training samples for corresponding individuals. The second column shows examples of individuals identified using fixed weighting. The third column shows examples of individuals identified after integrating AMPnet. Yellow boxes represent examples of corresponding individual training samples, red boxes represent identification failures, and green boxes represent successful identifications

From Fig. 10 Photo Intimacy Matching Module Results Analysis, we can observe that when there is a significant difference between training and testing samples, such as when there is incomplete facial information available as a reference, it is challenging to identify corresponding instances using fixed weighting. In these cases, the AMPnet feature fusion model can automatically determine which feature among facial features, head features, and body features contains more information to fuse, thereby improving our recognition accuracy. To comprehensively evaluate this model, this paper compared the fixed weighting method and the AMPnet feature fusion method using six commonly used metrics for evaluating multi-class models. The results are shown in Tables 4 and 5.

Table 4: Evaluation results on PIPA1.0

Metric	Method	Dataset			
		Day1.0	**Original1.0**	**Album1.0**	**Time1.0**
Weighted Precision*	Baseline	0.5054	0.7916	0.7376	0.6493
	+ AMPnet	**0.5627**	**0.8437**	**0.7917**	**0.7288**
Weighted Recall	Baseline	0.5110	0.8084	0.7364	0.6518
	+ AMPnet	**0.6348**	**0.8678**	**0.8267**	**0.7632**
Weighted F1-score	Baseline	0.4536	0.7651	0.6901	0.6051
	+ AMPnet	**0.5628**	**0.8365**	**0.7882**	**0.7124**
Macro Precision*	Baseline	0.3965	0.7132	0.6536	0.5793
	+ AMPnet	**0.3994**	**0.7841**	**0.7171**	**0.6269**
Macro Recall	Baseline	0.2864	0.6274	0.5285	0.4629
	+ AMPnet	**0.3776**	**0.7385**	**0.6761**	**0.5604**
Macro F1-score	Baseline	0.2941	0.6407	0.5531	0.4830
	+ AMPnet	**0.3526**	**0.7363**	**0.6715**	**0.5586**

Note: *Weighted precision is defined as the precision weighted by the proportion of instances per identity. Macro precision is computed as the arithmetic mean of per-class precision values. The bold data represent the best results.

Table 5: Evaluation results on PIPA2.0

Metric	Method	Dataset			
		Day2.0	**Original2.0**	**Album2.0**	**Time2.0**
Weighted precision	Baseline	0.4862	0.7872	0.7389	0.6732
	+ AMPnet	**0.5651**	**0.8104**	**0.7840**	**0.7142**
Weighted recall	Baseline	0.4966	0.8049	0.7336	0.6491
	+ AMPnet	**0.6276**	**0.8368**	**0.8102**	**0.7437**
Weighted F1-score	Baseline	0.4408	0.7665	0.6815	0.5771
	+ AMPnet	**0.5656**	**0.8033**	**0.7725**	**0.6984**
Macro precision	Baseline	0.3752	0.7267	0.6815	0.5771
	+ AMPnet	0.3693	0.7200	**0.7063**	**0.5820**
Macro recall	Baseline	0.2604	0.6530	0.5717	0.4413
	+ AMPnet	**0.3228**	**0.6717**	**0.6468**	**0.5147**
Macro F1-score	Baseline	0.2736	0.6655	0.5916	0.4681
	+ AMPnet	**0.3120**	**0.6707**	**0.6467**	**0.5135**

Note: The bold data represent the best results.

From Tables 4 and 5, we can see that the metrics after feature fusion with AMPnet are superior to those using a fixed weighting method. Additionally, this study found that when using the same method, the weighted-average metrics are significantly better than the corresponding macro-average metrics, especially evident in the day segmentation. Therefore, the study concludes that the macro-average metrics are affected by instances of certain classes with fewer and more challenging instances in the test set. However, overall, the method using feature fusion with AMPnet is generally superior to the fixed weighting method.

5.4.2 Photo Intimacy Matching Module Results Analysis

When the Visual Matching Module alone cannot accurately identify instances in the test set (i.e., when the maximum visual matching score of a particular instance is less than the threshold θ), this paper input the photos corresponding to the instance into the Photo Intimacy Matching Module for photo-level matching. This module seeks photos that have high similarity with the instance and can be identified solely based on visual clues (i.e., photos with high similarity and their corresponding instances have maximum visual matching scores greater than the threshold θ). This paper then use the photos with high intimacy obtained from the photo intimacy module to assist in identifying the instance. Fig. 11 shows some examples in the test set where the Visual Matching Module alone fails to identify correctly, but successful auxiliary identification is achieved after using the Photo Intimacy Matching Module.

Figure 11: Instances successfully identified by the photo intimacy matching module are grouped into three categories: (**a**) Represents instances where the Visual Matching Module alone fails to identify the corresponding instances in photos. (**c**) Represents photos successfully identified where the highest intimacy level is observed in the identified photo set. (**b**) Represents instances successfully assisted in identification by the photo with the highest intimacy level

Through these three sets of instances, we found that relying solely on vision makes it difficult to correctly identify the identities of independent instance photos that require assistance from the Photo Intimacy Matching Module. These instances may be challenging to identify accurately due to factors such as facing away from the camera, being too far from the camera, or significant differences in photo styles. Additionally, because these instances are independent, they cannot be identified through social relationships. Therefore, the Photo Intimacy Matching Module can find photos with higher similarity to the instances in accurately identified visual photos, thus assisting in identification. Unlike algorithms in previous studies that establish relationships between individuals in photos, the examples mentioned above include photos with only one instance. In such photos, the relationships between individuals do not provide any assistance in identifying instances, yet our method remains effective.

To better demonstrate the effectiveness of the Photo Intimacy Matching Module within the overall model, this paper analyzed the accuracy of the Photo Intimacy Matching Module's auxiliary identification, the accuracy of the Visual Matching Module alone, and the proportion of errors made solely by the Visual Matching Module in the total number of errors for the four splits of PIPA1.0 and PIPA2.0 with a threshold θ of 0.5. The results are shown in Fig. 12. We can see that at a threshold θ of 0.5, the majority of instances misidentified by relying solely on the Visual Matching Module are successfully identified with the assistance of the Photo Intimacy Matching Module. Additionally, at a threshold θ of 0.5, the accuracy of auxiliary identification using the Photo Intimacy Matching Module is higher across all four splits of the two datasets compared to relying solely on the Visual Matching Module. Therefore, we can conclude that when the Visual Matching Module alone fails to accurately identify the corresponding instances, incorporating the Photo Intimacy Matching Module for auxiliary identification can effectively improve the recognition accuracy in such cases.

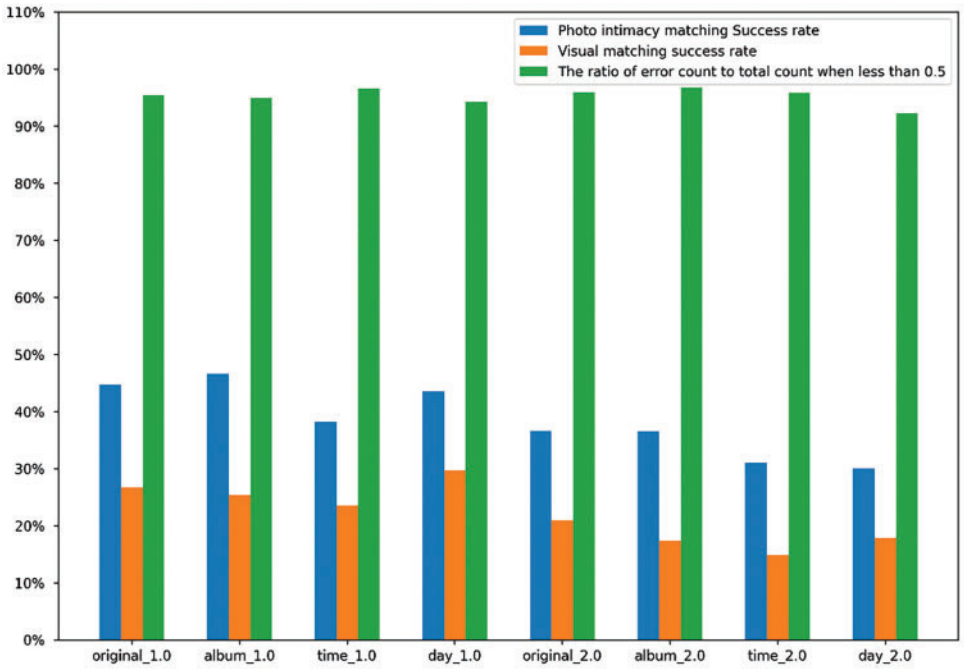

Figure 12: Photo intimacy matching module analysis figure

6 Conclusion and Future Research Direction

To address the issue of single-person-in-photos recognition that cannot be identified by visual or visual + social relationships alone in photo collections, this paper designed a new person recognition model. This model consists of two modules: the Visual Matching Module and the Photo Intimacy Matching Module. When the Visual Matching Module fails to accurately identify a specific person, the model uses the Photo Intimacy Matching Module for auxiliary identification, thus solving the recognition problem for independent instance photos. This paper conducted experiments on the publicly available PIPA1.0 and PIPA2.0 datasets, with DOI: 10.5281/zenodo.12508096 (accessed on 15 October 2025), to verify the effectiveness of the model. The results show that on the PIPA1.0 dataset, our model outperforms previous methods in three of the splits. On the PIPA2.0 dataset, our model also demonstrates strong effectiveness, with improvements across all four splits compared to the relevant baselines. Notably, in the most challenging day split, our Photo Intimacy Matching Module increased the accuracy from 62.76% to 69.70%.

Our analysis reveals three primary failure patterns: (1) Extreme pose variations—such as back-facing subjects with head yaw > 60° or pitch > 45°—where degraded visual features prevent reliable matching; (2) Heavy occlusions (>70% body area obscured) that deprive both visual and intimacy modules of discriminative cues, particularly in crowded scenes; (3) Low-intimacy contexts ($\max I(i, j) < 0.2$) where standalone photos lack correlated references for auxiliary recognition. Key limitations encompass: the static intimacy metric (fixed weights in Formula 6 constrain adaptability to diverse album structures), persistent cross-cultural recognition gaps (8.3% accuracy drop for Southeast Asian faces despite PIPA2.0 supplementation), and computational overhead ($O(n^2)$ operations requiring 42 h for 67 k images). Future work will address these constraints through dynamic intimacy weighting via graph networks and cross-cultural feature adaptation.

Although experiments have proven the effectiveness of the Photo Intimacy Matching Module, it still has certain limitations. Similar to previous studies' handling of different visual regions, the photo intimacy matrix is calculated using a fixed weighting method. For future work, we plan to use methods such as graph neural networks to obtain the photo intimacy matrix, allowing it to dynamically measure the intimacy relationship between photos in different situations. Although this study focuses on unimodal analysis, it can be extended to multimodal scenarios in the future to support multimedia applications.

Acknowledgement: The authors would like to acknowledge the publicly available PIPA dataset which served as the foundation of our work.

Funding Statement: This research was supported by "the Fundamental Research Funds for the Central Universities" (Grant Nos.: 3282025045, 3282024008), "Science and Technology Project of the State Archives Administration of China" (Grant No.: 2025-Z-009).

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization, Xiuying Li and Xiaoyi Duan; methodology, Xiaoyi Duan and Tianqi Zou; software, Tianqi Zou, Chenyang Wang and Yu Gu; validation, Chenyang Wang; writing—original draft preparation, Xiaoyi Duan; writing—review and editing, Xiuying Li; visualization, Chenyang Wang; supervision, Tianqi Zou. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: Data openly available in a public repository.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Sheng B, Li P, Zhang Y, Wang X, Yang C. GreenSea: visual soccer analysis using broad learning system. IEEE Trans Cybern. 2021;51(3):1463–77. doi:10.1109/tcyb.2020.2988792.
2. Zeghoud S, Ali SG, Ertugrul E, Demirci MF, Tadjine A. Real-time spatial normalization for dynamic gesture classification. Vis Comput. 2022;38(7):2209–24.
3. Ali SG, Wang X, Li P, Yang X, Zhang Y. A systematic review: virtual-reality-based techniques for human exercises and health improvement. Front Public Health. 2023;11:1143947. doi:10.3389/fpubh.2023.1143947.
4. Hassan MM, Hussein HI, Eesa AS, Abdalla AH, Ali HS. Face recognition based on gabor feature extraction followed by FastICA and LDA. Comput Mater Contin. 2021;68(2):1637–59. doi:10.32604/cmc.2021.016467.
5. Kim M, Jain AK, Liu X. AdaFace: quality adaptive margin for face recognition. In: Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 18–24; New Orleans, LA, USA. p. 18750–9.

6. Park H, Ham B. Relation network for person re-identification. AAAI Conf Artif Intell. 2020;34(7):11839–47. doi:10.1609/aaai.v34i07.6857.
7. Wan F, Wu Y, Qian X, Liu Y, Wan J. When person re-identification meets changing clothes. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); 2020 Jun 14–19; Seattle, WA, USA. p. 830–1.
8. Liu Y, Chen J, Li Y, Zhou H. Joint face normalization and representation learning for face recognition. Pattern Anal Appl. 2024;27:64.
9. III DC, Brogan J, Barber N, Aykac D, Baird S, Burchfield N, et al. Expanding accurate person recognition to new altitudes and ranges: the BRIAR dataset. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops; 2023. p. 593–602.
10. Viola M, Voto C. Designed to abuse? Deepfakes and the non-consensual diffusion of intimate images. Synthese. 2023;201(30):1–20. doi:10.1007/s11229-022-04012-2.
11. Saad RSM, Moussa MM, Abdel-Kader NS, El-Batt T. Deep video-based person re-identification (Deep Vid-ReID): comprehensive survey. EURASIP J Adv Signal Process. 2024;2024:63. doi:10.1186/s13634-024-01139-x.
12. Mahajan A, Singla SK. DeepBio: a deep CNN and Bi-LSTM learning for person identification using ear biometrics. Comput Model Eng Sci. 2024;141(2):1623–49. doi:10.32604/cmes.2024.054468.
13. Zhang N, Paluri M, Taigman Y, Lin D, Duerig T. Beyond frontal faces: improving person recognition using multiple cues. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2015 Jun 7–12; Boston, MA, USA. p. 4804–13.
14. Wang Z, Chen T, Ren J, Wen J, Sun X. Deep reasoning with knowledge graph for social relationship understanding. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI); 2018 Jul 13–19; Stockholm, Sweden. p. 1021–8.
15. Huang Q, Liu W, Lin D. Person search in videos with one portrait through visual and temporal links. In: Computer vision—ECCV 2018. vol. 11205 of Lecture Notes in Computer Science. Cham, Switzerland: Springer; 2018. p. 437–54.
16. Goel A, Ma KT, Tan C. An end-to-end network for generating social relationship graphs. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019 Jun 15–20; Long Beach, CA, USA. p. 11186–95.
17. Li W, Duan Y, Lu J, Feng J, Zhou J. Graph-based social relation reasoning. In: Computer vision—ECCV 2020. vol. 12360 of Lecture Notes in Computer Science. Cham, Switzerland: Springer; 2020. p. 18–34.
18. Li H, Brandt J, Lin Z, Bourdev L. A multi-level contextual model for person recognition in photo albums. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016 Jun 27–30. Las Vegas, NV, USA. p. 1297–305.
19. Li Y, Lin G, Zhuang B, Shen C. Sequential person recognition in photo albums with a recurrent network. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 Jul 21–26; Honolulu, HI, USA. p. 5660–8.
20. Xia J, Rao A, Huang Q, Lin D, Wei Q. Online multi-modal person search in videos. In: Computer vision—ECCV 2020. vol. 12374 of Lecture Notes in Computer Science. Cham, Switzerland: Springer; 2020. p. 174–90.
21. Xu T, Zhou P, Hu L, Huang Q, Lin D. Socializing the videos: a multimodal approach for social relation recognition. ACM Trans Multimed Comput, Commun Appl. 2021;17(1):1–23. doi:10.1145/3416493.
22. Brown A, Kalogeiton V, Zisserman A. Face, body, voice: video person-clustering with multiple modalities. In: Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW); 2021 Oct 11–17. Montreal, QC, Canada. p. 3184–94.
23. Oh SJ, Benenson R, Fritz M, Schiele B. Person recognition in personal photo collections. In: Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV); 2015 Dec 7–13; Santiago, Chile. p. 3862–70.
24. Kumar V, Namboodiri A, Paluri M, Jain AK. Pose-aware person recognition. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 Jul 21–26; Honolulu, HI, USA. p. 6223–32.

25. Xue J, Meng Z, Katipally K, Cai J, Prasanna P. Clothing change aware person identification. In: Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); 2018 Jun 18–22; Salt Lake City, UT, USA. p. 2112–20.
26. Huang Q, Xiong Y, Lin D. Unifying identification and context learning for person recognition. In: Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2018 Jun 18–22; Salt Lake City, UT, USA. p. 2217–25.
27. Sousa EV, Macharet DG. Structural reasoning for image-based social relation recognition. Comput Vis Image Underst. 2023;235(3):103785. doi:10.1016/j.cviu.2023.103785.
28. Li P, Sheng B, Chen CLP. Face sketch synthesis using regularized broad learning system. IEEE Trans Neural Netw Learn Syst. 2021;33(10):5346–60. doi:10.1109/tnnls.2021.3070463.
29. Wang Y, Zhang C, Liao X, Wang X, Gu Z. An adversarial attack system for face recognition. J Artif Intell. 2021;3(1):1–8. doi:10.32604/jai.2021.014175.
30. Khoshnevisan E, Hassanpour H, AlyanNezhadi M. Face recognition based on general structure and angular face elements. Multimed Tools Appl. 2024;83(36):83709–27. doi:10.1007/s11042-024-18897-3.
31. Jing HR, Lin GJ, Chen TT, Zhang HJ, Zhang L, Zhou SY. Unrestricted face recognition algorithm based on improved residual network IR-ResNet-SE. J Comput. 2023;34(2):29–39.
32. Karambakhsh A, Kamel A, Sheng B, Li P, Yang P, Feng DD. Deep gesture interaction for augmented anatomy learning. Int J Inf Manag. 2019;45(4):328–36. doi:10.1016/j.ijinfomgt.2018.03.004.
33. Kamencay P, Benco M, Mizdos T, Radil R. A new method for face recognition using convolutional neural network. Adv Electr Electron Eng. 2017;15(4):663–72. doi:10.15598/aeee.v15i4.2389.
34. Coskun M, Ucar A, Yildirim o, Demir Y. Face recognition based on convolutional neural network. In: 2017 International Conference on Modern Electrical and Energy Systems (MEES); 2017 Nov 15–17. Kremenchuk, Ukraine. p. 376–9.
35. Wang J, Li Z. Research on face recognition based on CNN. IOP Conf Series Earth Environ Sci. 2018;170:032110. doi:10.1088/1755-1315/170/3/032110.
36. Wang D, Yu H, Wang D, Li G. Face recognition system based on CNN. In: 2020 International Conference on Computer Information and Big Data Applications (CIBDA); 2020 Apr 17–19; Guiyang, China. p. 470–3.
37. Khan S, Naseer M, Hayat M, Zamir SW, Khan FS, Shah M. Transformers in vision: a survey. ACM Computi Surv (CSUR). 2022;54(10s):1–41. doi:10.1145/3505244.
38. Zhou D, Kang B, Jin X, Yang L, Lian X, Jiang Z, et al. Deepvit: towards deeper vision transformer. arXiv:2103.11886. 2021.
39. Han K, Wang Y, Chen H, Chen X, Guo J, Liu Z, et al. A survey on vision transformer. IEEE Trans Pattern Anal Mach Intell. 2022;45(1):87–110. doi:10.1109/tpami.2022.3152247.
40. Zhong Y, Deng W. Face transformer for recognition. arXiv:2103.14803. 2021.
41. Dan J, Liu Y, Xie H, Deng J, Xie H, Xie X, et al. TransFace: calibrating transformer training for face recognition from a data-centric perspective. In: Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision; 2023 Oct 1–6; Paris, France. p. 20642–53.
42. Opanasenko VM, Fazilov SK, Mirzaev ON, Kakharov SSU. An ensemble approach to face recognition in access control systems. J Mob Multimed. 2024;20(3):749–68. doi:10.13052/jmm1550-4646.20310.
43. Ding Z, Zhang X, Xia Z, Jebe L, Tu Z, Zhang X. DiffusionRig: learning personalized priors for facial appearance editing. In: Proceedings of the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2025 Jun 10–17; Nashville, TN, USA. p. 12736–46.
44. Zheng L, Yang Y, Hauptmann AG. Person re-identification: past, present and future. arXiv:1610.02984. 2016.
45. He S, Luo H, Wang P, Wang F, Li H, Jiang W. Transreid: transformer-based object re-identification. In: Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision; 2021 Oct 10–17; Montreal, QC, Canada. p. 15013–22.
46. Li Y, He J, Zhang T, Liu X, Zhang Y, Wu F. Diverse part discovery: occluded person re-identification with part-aware transformer. In: Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2021 Jun 20–25; Nashville, TN, USA. p. 2898–907.

47. Chen Y, Xia S, Zhao J, Zhou Y, Niu Q, Yao R, et al. ResT-ReID: transformer block-based residual learning for person re-identification. Pattern Recognit Lett. 2022;157(8):90–6. doi:10.1016/j.patrec.2022.03.020.
48. Bai N, Wang X, Han R, Wang Q, Liu Z. PAFormer: anomaly detection of time series with parallel-attention transformer. IEEE Trans Neural Netw Learn Syst. 2025;36(2):3315–28. doi:10.1109/tnnls.2023.3337876.
49. Zheng L, Shen L, Tian L, Wang S, Wang J, Tian Q. Scalable person re-identification: a benchmark. In: Proceedings of the 2015 IEEE International Conference on Computer Vision; 2015 Dec 7–13; Santiago, Chile. p. 1116–24.