Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.060783

ARTICLE

VPM-Net: Person Re-ID Network Based on Visual Prompt Technology and Multi-Instance Negative Pooling

Haitao Xie, Yuliang Chen, Yunjie Zeng, Lingyu Yan, Zhizhi Wang and Zhiwei Ye*

School of Computer Science, Hubei University of Technology, Wuhan, 430068, China

*Corresponding Author: Zhiwei Ye. Email: weizhiye121@163.com

Received: 09 November 2024; Accepted: 17 February 2025; Published: 16 April 2025

ABSTRACT: With the rapid development of intelligent video surveillance technology, pedestrian re-identification has become increasingly important in multi-camera surveillance systems. This technology plays a critical role in enhancing public safety. However, traditional methods typically process images and text separately, applying upstream models directly to downstream tasks. This approach significantly increases the complexity of model training and computational costs. Furthermore, the common class imbalance in existing training datasets limits model performance improvement. To address these challenges, we propose an innovative framework named Person Re-ID Network Based on Visual Prompt Technology and Multi-Instance Negative Pooling (VPM-Net). First, we incorporate the Contrastive Language-Image Pre-training (CLIP) pre-trained model to accurately map visual and textual features into a unified embedding space, effectively mitigating inconsistencies in data distribution and the training process. To enhance model adaptability and generalization, we introduce an efficient and task-specific Visual Prompt Tuning (VPT) technique, which improves the model's relevance to specific tasks. Additionally, we design two key modules: the Knowledge-Aware Network (KAN) and the Multi-Instance Negative Pooling (MINP) module. The KAN module significantly enhances the model's understanding of complex scenarios through deep contextual semantic modeling. MINP module handles samples, effectively improving the model's ability to distinguish fine-grained features. The experimental outcomes across diverse datasets underscore the remarkable performance of VPM-Net. These results vividly demonstrate the unique advantages and robust reliability of VPM-Net in fine-grained retrieval tasks.

KEYWORDS: Person re-identification; multi-instance negative pooling; visual prompt tuning

1 Introduction

The Person Re-Identification (ReID) task aims to efficiently retrieve matching images from large-scale databases based on textual descriptions or key attributes [1–3]. Traditional methods typically rely on simple text-based search queries, but these approaches have limitations in query efficiency and accuracy. As a result, research has gradually shifted toward cross-modal pedestrian retrieval techniques that integrate both text and visual modalities [4–6]. Compared to image-based retrieval, textual descriptions offer more flexibility in summarizing individual characteristics, thereby improving the usability and convenience of retrieval systems. Text-based retrieval of pedestrian images is highly significant in various application scenarios, including personal photo management and high-risk domains such as public safety and national security.

However, despite the growing importance of this task, it still faces several challenges, particularly in meeting real-time processing requirements. First, pedestrian retrieval must be completed quickly on large datasets, placing high demands on computational speed. Existing methods often rely on complex neural

network architectures, resulting in long inference times and difficulties in meeting real-time requirements. Second, the inherent heterogeneity in cross-modal representations [7] exacerbates this issue, leading to discrepancies between text and images. Text descriptions may introduce ambiguities related to word order and human interpretation, increasing the complexity of model learning, especially when dealing with large numbers of negative samples in complex scenarios. Finally, the class imbalance problem further intensifies these challenges. In real-world applications, only a small portion of images contain relevant information that can effectively improve model performance, making pedestrian recognition in complex environments even more difficult.

To address these challenges, previous research has primarily focused on optimizing representation learning and developing efficient cross-modal matching strategies between text and images. Early global matching methods [8,9] attempted to map images and text to a shared embedding space. However, these methods often applied loss functions only at the network output layer, neglecting the importance of intermediate modality interactions. This limitation prompted the development of local alignment modules [10–13] aimed at achieving more fine-grained matching. However, these methods may introduce noise and uncertainty, requiring the extraction and storage of multiple local image-text matching pairs, and conducting time-consuming similarity calculations, further increasing the complexity of model maintenance and expansion. Furthermore, current mainstream methods typically require reconfiguring the entire backbone network architecture each time adjustments are made, leading to significant time overhead and low retrieval efficiency. While these methods have made considerable progress in retrieval performance, their adaptability is limited, especially in complex environments that require dynamic adjustments, making them inadequate for real-world applications. Therefore, there is an urgent need for innovative solutions that effectively balance performance and adaptability to meet the evolving demands of cross-modal retrieval tasks.

To address the multiple challenges in text-to-image retrieval tasks, this paper proposes a balanced and efficient innovative framework VPM-Net. This framework introduces three modules: Visual Prompt Tuning (VPT) [14], Knowledge-Aware Network (KAN), and MINP. The goal is to significantly enhance the performance of text and image encoders, particularly in handling cross-modal alignment and class imbalance issues. Firstly, VPM-Net adopts the CLIP architecture [15] as the backbone, combined with the latest VPT strategy. Unlike traditional visual prompt methods, the VPT module does not freeze the core parts of the backbone model but selectively freezes the parameters of the tail layers. While this strategy may not yield the same significant performance improvements as freezing the core, it provides stable performance enhancement by avoiding excessive limitations on the model's flexibility. This greatly improves the adaptability of the model in the process of rapid fine-tuning. To further enhance the model's understanding of multimodal information, the KAN module is introduced. The KAN module focuses on improving the acquisition and utilization of contextual information between the visual and textual modalities. It integrates both linear and nonlinear knowledge representations, significantly boosting feature understanding and modality alignment capabilities. By optimizing the information exchange between modalities, the KAN module helps the model capture the subtle differences between text descriptions and visual information, thus improving overall recognition accuracy. Finally, the MINP module is designed to enhance the model's ability to learn from difficult samples. Traditional training methods often neglect the impact of challenging samples, leading to models being easily distracted by easy samples, thus limiting performance. The MINP module targets difficult samples, encouraging the model to become more sensitive to these challenging examples. This module not only optimizes the model's local alignment capability but also enhances its robustness in complex environments.

VPM-Net effectively solves multiple challenges in text-to-image retrieval tasks by innovatively combining the VPT, KAN, and MINP modules. The VPT module improves the model's adaptability and flexibility, the KAN module enhances contextual alignment between modalities, and the MINP module boosts the model's discriminative ability by focusing on difficult samples. Through extensive experimental validation, VPM-Net demonstrates significant advantages over other popular methods on multiple benchmark datasets. The primary contributions of this paper are summarized as follows:

- This study represents the first application of visual prompt technology within the ReID cross-modal domain, introducing the innovative VPT module. The findings demonstrate that VPT can enhance performance primarily through parameter quantity, though it is not suitable for reverse applications.
- We introduce the KAN module, which enriches the model's understanding of contextual information from both visual and textual modalities, enhancing feature representation and improving the overall discriminative power of the model.
- The paper presents the MINP module, which refines the model's focus on challenging samples while disentangling them from a surplus of easy negative samples. This refinement facilitates the extraction of valuable yet submerged samples, thereby enhancing learning and boosting model accuracy.
- The research findings confirm that our proposed state-of-the-art approach excels in multiple facets,successfully implemented on the Chinese University of Hong Kong Person Description Dataset (CUHK-PEDES) [16] and Image-Caption Fine-grained Pedestrian Description Evaluation Set (ICFG-PEDES) [11] datasets, surpassing the performance of baseline methods and Implicit Relation Reasoning and Aligning for Text-to-Image Person Retrieval (IRRA) [17] across three common datasets.

2 Related Work

The ReID (Re-Identification) task, as exemplified by the CUHK-PEDES dataset [16], has consistently posed a critical challenge in aligning and matching image and text features within a unified embedding space. This section reviews foundational approaches in the field, highlighting the evolution of techniques that aim to solve the complex alignment problems in cross-modal retrieval. The discussion proceeds from early feature extraction and alignment strategies to advanced methods leveraging prompt tuning in vision-language tasks, culminating in our proposed modules to enhance cross-modal alignment accuracy and robustness.

2.1 Early Approaches in ReID and the Evolution of Cross-Modal Alignment Techniques

Initial approaches to the ReID problem primarily focused on leveraging deep learning architectures to extract and align features from both visual and textual data, attempting to bring them into a shared representation space. These methods often employed networks such as Very deep convolutional networks for large-scale image recognition (VGG) [18] to extract visual representations, paired with Long short-term memory (LSTM) [19] for textual encoding [16,20,21]. Such models were typically trained with matching loss functions, which were designed to align the modalities by minimizing the distance between matched image-text pairs in the embedding space [22]. However, these early models faced challenges in capturing finer-grained details necessary for accurate retrieval, often struggling with the nuances required for real-world ReID tasks.

With advancements in natural language processing (NLP), more sophisticated techniques were introduced, marking a shift in the complexity and capability of feature extraction and alignment. One notable advancement was the emergence of the cFine model [23], which integrated the powerful BERT [24] transformer model within a ResNet framework. This combination allowed for the joint extraction of both image and text features while introducing cross-modal matching loss functions that facilitated global alignment within a unified embedding space. Such methods not only improved alignment quality but also laid the foundation for more robust cross-modal retrieval.

Other research groups focused on enhancing these alignment models by incorporating local feature learning modules. For instance, explicit attribute-based models were developed to capture key characteristics such as age and gender, essential for person re-identification [10,13,25–27]. Meanwhile, methods utilizing attention mechanisms emerged to implicitly learn local features that improved the model's adaptability and precision [11,28–30]. A novel space-cover convolution method is proposed to learn local high-order features by constructing an anisotropic spatial geometry in feature space, which enables the capture of implicit shape representations [31]. The recently proposed multi-view images complement the detailed features of 3D objects by adjusting the rendering viewpoint, helping to understand people more fully in both holistic and occluded ReID situations [32]. Although these approaches contributed to substantial performance improvements, they often introduced added complexity through the need for additional supervision signals, which could lead to noise and overfitting issues, potentially compromising generalization across datasets.

A significant breakthrough in the field came with the introduction of CLIP (Contrastive Language-Image Pretraining) by Yan et al. [23], designed specifically for text-to-image character retrieval. CLIP represented a shift towards implicit alignment techniques, aligning image and text features within a single model architecture by training on extensive paired data. The development of CLIP spurred interest in implicit alignment strategies, such as the IRRA model [17], which extended these concepts. Methods within this paradigm either focused on explicit alignment of independently pre-trained modalities or achieved implicit alignment within the CLIP framework itself [23,33]. Despite these improvements in mapping quality and rank-1 accuracy, a persistent challenge remained: differentiating between easy and hard samples. Existing alignment techniques struggled to handle this, resulting in suboptimal performance on more complex instances.

To address these limitations, we introduce the MINP module, specifically designed to tackle the challenges associated with class imbalance and complex alignment. Across existing datasets, class imbalance remains a prevalent issue, with a disproportionately small subset of samples matching the target accurately. Traditional loss functions, such as cross-entropy and its variants, are typically optimized for direct alignment discrepancies between modalities but lack mechanisms to address class imbalance. Inspired by the focal loss [34], our proposed MINP Loss incorporates a modulation factor, which effectively addresses the imbalance by weighting challenging samples more heavily, thereby enhancing the model's representation capacity for underrepresented instances. The components of this module and its integration into our architecture are further detailed in Section 3.2.

2.2 Prompt Tuning in Vision-Language Tasks and the Proposal of VPT-Reverse and VPT-Tail

Recent advancements in NLP have introduced the "Pre-train, Prompt, Predict" paradigm, an alternative to the conventional pre-training and fine-tuning approach. Researchers such as Jia et al. [14] demonstrated that by introducing prompt-based methods, models could better accommodate downstream tasks without the need for extensive task-specific tuning. This paradigm shift allows for a more flexible integration of pre-trained language models, which can be adapted to a wider range of applications with minimal adjustments.

Extending this concept to the computer vision domain, the VPT approach, presented at ECCV 2022 [14], pioneered the application of prompt tuning for vision-language tasks. The VPT method introduced two variants, VPT-Shallow and VPT-Deep, which showed that model performance could be significantly enhanced with a modest increase in parameters. This approach allowed for efficient adaptation of large models to vision-language tasks without the need for full fine-tuning, preserving computational efficiency.

Transformer models can acquire extensive linguistic knowledge and deep semantic information.Building on this foundation, we propose two novel variants, VPT-Portion and VPT-Tail, specifically designed to improve prompt flexibility within transformer architectures. VPT-Portion and VPT-Tail introduce enhanced prompt structures that enable finer control over the tuning process. In particular, VPT-Tail extends the utility of prompt tuning by focusing on fine-grained local alignment, an area critical for cross-modal retrieval. When combined with the proposed MINP Loss, VPT-Tail not only improves overall model effectiveness but also achieves superior performance in alignment tasks involving complex, localized features. This combination of techniques offers a comprehensive and efficient solution to address the challenges of cross-modal retrieval, particularly in vision-language scenarios.

3 Method

In this section, we present the proposed VPM-Net framework, a novel approach aimed at enhancing cross-modal retrieval through innovative prompt tuning and loss modulation strategies. Fig. 1 provides a comprehensive overview of the VPM-Net architecture, showcasing how each component integrates to achieve robust alignment and retrieval between image and text modalities. The VPM-Net framework is designed to address key challenges in cross-modal feature alignment by combining advanced prompt tuning with optimized loss functions, allowing it to capture both global and local feature alignments. The architecture of VPM-Net is structured around three primary components: (1) the **Prompt Module**, which utilizes advanced visual prompt tuning strategies to enhance flexibility and fine-grained feature alignment; (2) the **KAN Module**, a knowledge-aware component that captures contextual nuances in multimodal data, improving the adaptability of the model to varying semantic elements; and (3) the **Matching and Loss Computation Modules**, which employ a novel loss function to address class imbalance and refine alignment by prioritizing challenging samples. Together, these components work to effectively map image and text features within a shared embedding space, ensuring robust cross-modal alignment and optimizing retrieval accuracy across a range of tasks. In the following subsections, each of these components is explained in detail, highlighting their roles and contributions to the overall performance of the model.

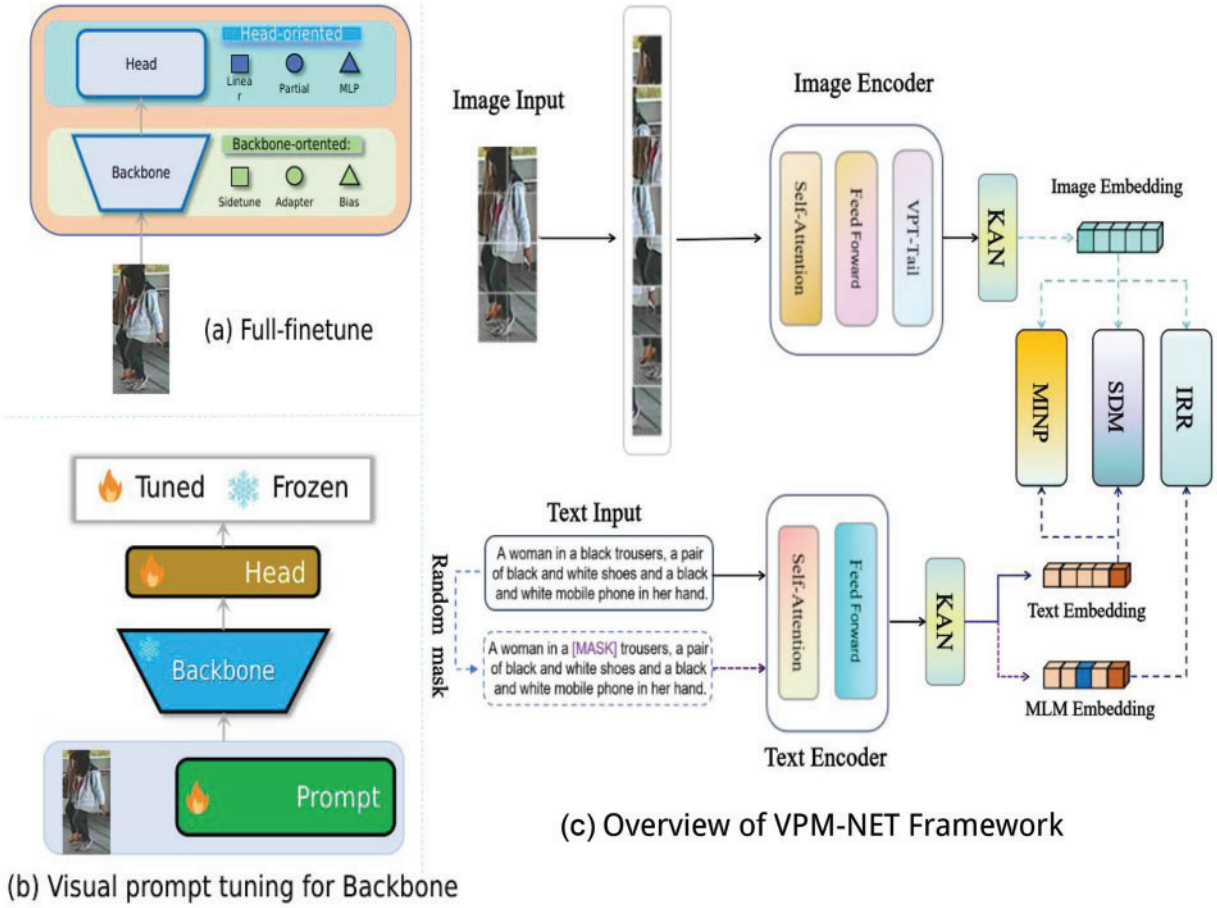

Figure 1: Overview of the proposed framework. (a) In existing person re-identification methods, the primary model adaptation strategy employed during model tuning is full fine-tuning. (b) VPT avoids fine-tuning the Transformer itself. Instead, it introduces a small set of task-specific, learnable parameters at the input layer while freezing the Transformer backbone to reduce the number of tunable parameters. (c) The VPM-Net framework enhances cross-modal retrieval through visual prompt tuning and loss modulation. During model adaptation, only a small number of additional parameters are adjusted to fine-tune the entire re-identification (ReID) model. The framework integrates a KAN module during representation learning to capture contextual nuances and improve adaptability. Additionally, a MINP module is introduced in metric learning to refine alignment and improve retrieval accuracy

3.1 Visual Prompt Tuning (VPT)

The Person Re-Identification (ReID) task requires rapid retrieval from large-scale datasets, placing extremely high demands on computational speed. Existing models often improve performance by increasing complexity [9,35,36], but this approach typically leads to inefficiency and limited scalability in real-world applications. To address this issue, we introduce the Visual Prompt Tuning (VPT) module, which ensures high performance while avoiding excessive freezing of the model's core components. Unlike traditional methods, the VPT module selectively freezes the parameters at the tail end rather than freezing the entire backbone network. While this strategy does not achieve the same dramatic performance gains as freezing the core, it provides stable improvements in performance by avoiding over-restriction of the model's flexibility. This approach significantly enhances model adaptability, especially during rapid fine-tuning. As shown in Fig. 2, the VPT module is a key component of the proposed VPM-Net framework, designed to improve the efficiency and flexibility of cross-modal alignment. Additionally, we have designed two customized fast

adjustment strategies: VPT-Tail and VPT-Portion, which optimize specific layers within the Transformer network, reducing model complexity while focusing on achieving fine-grained alignment necessary for effective image-text retrieval [37].

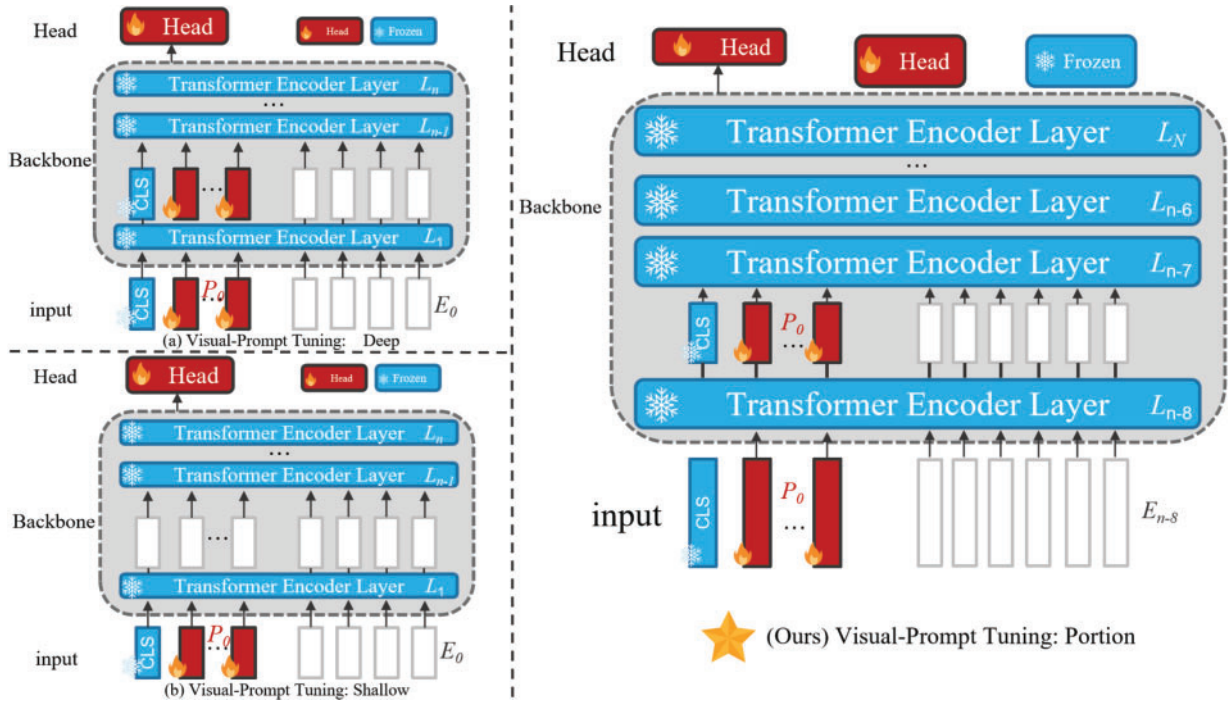

Figure 2: VPT network diagram in VPM-Net. In prior research, conventional methods have predominantly involved freezing all layers within the Transformer architecture and subsequently incorporating prompt vectors commencing from the initial layer. (a) VPT-Deep introduces prompt vectors at every layer throughout the entire architecture. (b) VPT-Shallow inserts prompt vectors exclusively at the first layer (Ours). VPT-Portion selectively freezes a subset of layers within the Transformer architecture and adopts a more adaptive strategy for prompt vector insertion. Specifically, prompt vectors are strategically inserted into the intermediate layers of the Transformer. This design philosophy is underpinned by the objective of attaining an optimal equilibrium between prompt efficiency and model stability, thereby enhancing the overall robustness and effectiveness of the model

Image Encoder: To extract robust image features, we employ a Vision Transformer (ViT) pre-trained with CLIP, taking an input image $I \in \mathbb{R}^{H\times W\times C}$ and dividing it into non-overlapping patches. This division results in $N = \frac{H\times W}{P^2}$ patches, each of size $P \times P$, which are linearly projected into a sequence of 1D tokens $\{f_i^v\}_{i=1}^N$. We then add positional embeddings and a [CLS] token to create the sequence $\{f_{\text{cls}}^v, f_1^v, \ldots, f_N^v\}$. This sequence is processed by L Transformer layers, capturing the spatial relationships among patches. A final linear projection maps f_{cls}^v to a shared embedding space for alignment with the text embeddings, providing a comprehensive representation of the image.

VPT-Tail approach introduces prompts only in the last few Transformer layers, focusing on enhancing representation at the output. By applying learnable prompt vectors to the final four layers (denoted as I_{n-4} to I_n), VPT-Tail ensures the Transformer captures nuanced features critical for high-quality alignment, without burdening the entire model with additional parameters. Each prompt token is a vector in $\mathbb{R}^d$, represented by the set $P = \{p^k \in \mathbb{R}^d \mid k = 1, \ldots, p\}$. The selective placement of prompts enables the model to effectively refine representations in the final stages, thus enhancing the retrieval accuracy and interpretability of the

model with minimal complexity. This VPT-Tail tuning process is formulated as:

$$\{x_n, Z_n, E_n\} = L_n([x_{n-1}, Z_{n-1}, E_{n-1}]), \quad n = 1, \ldots, N-4 \tag{1}$$

$$\{x_i, Z_i, E_i\} = L_i([x_{i-1}, P_{i-1}, E_{i-1}]), \quad i = N-3, \ldots, N \tag{2}$$

$$y = \text{Head}(X_N) \tag{3}$$

Text Encoder: The text embeddings are generated using the CLIP Transformer-based text encoder [38]. The encoder tokenizes the input text T via byte pair encoding (BPE) with a vocabulary of 49,152 tokens [39], converting text into a structured sequence that facilitates effective cross-modal alignment. This encoding process breaks down input text into interpretable units, enriching the representation for subsequent integration with image embeddings.

VPT-Portion: Complementing VPT-Tail, our **VPT-Portion** strategy introduces prompts selectively to the middle-to-end layers, ranging from L_{n-8} to L_n, based on experimental findings that this configuration enhances performance. By placing prompts at these targeted layers, VPT-Portion supports the model in capturing intermediate features that contribute to a more comprehensive representation of complex visual and textual elements. Each prompt token in this configuration is also a vector in $\mathbb{R}^d$, where $P = \{p^k \in \mathbb{R}^d \mid m - 8 \leq k \leq m\}$. This layer-specific prompt tuning enables the model to balance expressiveness and efficiency, focusing computational resources where they have the most impact.

The VPT-Portion process is defined as:

$$\{x_n, Z_n, E_n\} = L_n([x_{n-1}, Z_{n-1}, E_{n-1}]), \quad n = 1, \ldots, N-8 \tag{4}$$

$$\{x_i, Z_i, E_i\} = L_i([x_{i-1}, P_{i-1}, E_{i-1}]), \quad i = N-7, \ldots, N \tag{5}$$

$$y = \text{Head}(X_N) \tag{6}$$

By precisely adjusting the VPT-Tail and VPT-Portion strategies, the model can selectively leverage customized prompts to enhance the capabilities of the Transformer layers, thereby improving the performance of cross-modal alignment tasks. When VPT-Tail is introduced, the model adds 0.077 M parameters; whereas with VPT-Portion, the additional parameters amount to 0.923 M. These extra parameters represent a very small proportion of the total parameters in the ViT-base model, accounting for only 0.08% and 1.06%, respectively. This fine-tuning strategy not only enhances the model's adaptability but also minimizes unnecessary complexity, achieving an optimal balance between performance and efficiency.

3.2 KAN Module

To compensate for the inadequate performance of the existing MLP model in handling both linear and nonlinear relationships, we employ the KAN module (As shown in Fig. 3) to effectively unify the representation of these two types of relationships. At the level of representation learning, this module enhances the stability of model adjustments. we introduce the KAN (Knowledge-Aware Network) [40] module to enhance performance in the Text-to-Image Person Retrieval task. KAN is inspired by the Kolmogorov-Arnold theorem, which states that any function $f(x) = f(x_1, \ldots, x_n)$ can be represented as:

$$f(x) = \sum_{q=1}^{2n+1} \phi_q \left(\sum_{p=1}^{n} \phi_{q,p}(x_p) \right) \tag{7}$$

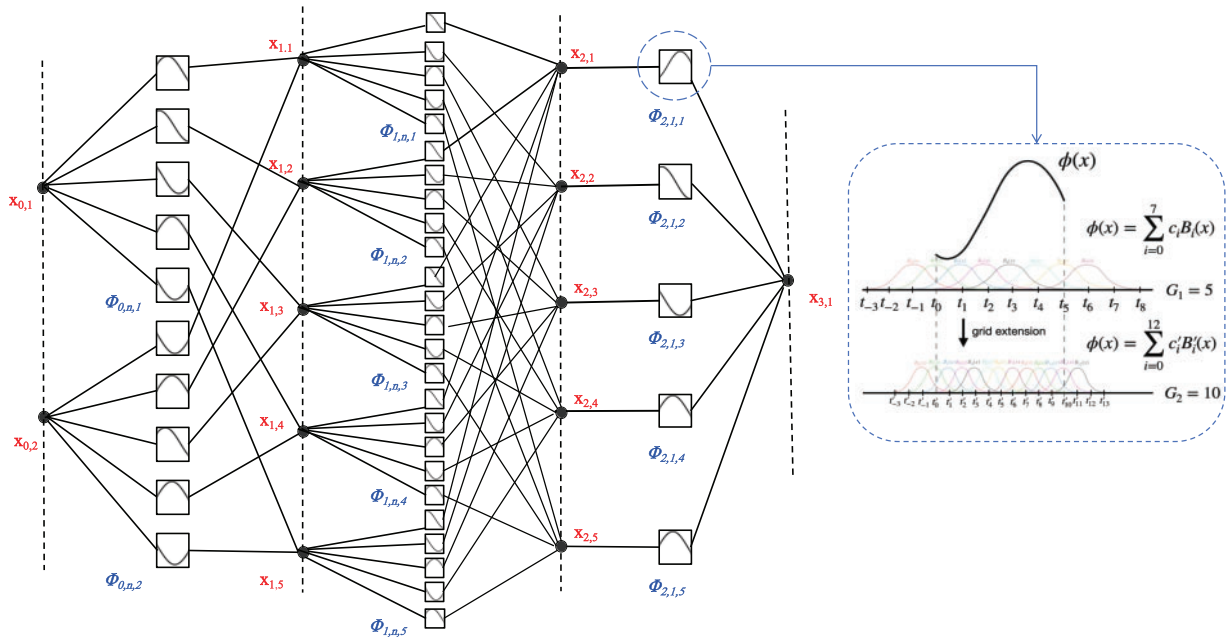

Figure 3: KAN network diagram in VPM-Net. We integrated KAN layers into text-to-image retrieval, reducing reliance on linear matrices and enabling direct learning of nonlinear activation functions

This theorem suggests that complex high-dimensional functions can be expressed in terms of simpler metafunctions, implying that learning high-dimensional functions can reduce learning one-dimensional functions. However, the diversity and complexity of these one-dimensional functions present significant challenges in practical applications. The KAN module extends the architecture of traditional neural networks, allowing for arbitrary width and depth, based on the premise that Eq. (7) involves only two nonlinearities and a limited number of terms $(2n + 1)$. Within this framework, the activation values for the$L$-layer KAN are represented as $\phi_{l,j,i}$, with the previous activation denoted by $x_{l,j,i} \approx \phi_{l,j,i}(x_{l,i})$. This transformation can be expressed in matrix form as follows:

$$X_{l+1} = \begin{pmatrix} \phi_{l,1,1}(\cdot) & \phi_{l,1,2}(\cdot) & \cdots & \phi_{l,1,n_l}(\cdot) \\ \phi_{l,2,1}(\cdot) & \phi_{l,2,2}(\cdot) & \cdots & \phi_{l,2,n_l}(\cdot) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_{l,L+1,1}(\cdot) & \phi_{l,L+1,2}(\cdot) & \cdots & \phi_{l,L+1,n_l}(\cdot) \end{pmatrix} X_l \tag{8}$$

Given an input vector $x_0 \in \mathbb{R}^{n_0}$, the output of KAN can be defined as:

$$KAN(X) = (\phi_{L-1} \circ \phi_{L-2} \circ \ldots \circ \phi_1 \circ \phi_0)X \tag{9}$$

Before the emergence of KAN, multilayer perceptrons (MLPs) were widely utilized in neural network architectures. While MLPs, including transformer models, rely on linear combinations and activation functions, this dependency inherently limits their ability to represent complex functions. This limitation can be mathematically expressed as:

$$MLP(X) = (W_{L-1} \circ \sigma \circ W_{L-2} \circ \sigma \circ \cdots \circ W_0)X \tag{10}$$

Here, W denotes the linear transformations, and σ represents the nonlinear activation functions. In contrast, KAN consolidates these transformations within a unified framework:

$$KAN(X) = \Phi(X) \tag{11}$$

Although KAN offers significant advantages, its practical application is limited by specific domain constraints. This is mainly due to its novel architecture, which may cause performance instability in certain scenarios. The U-KAN method [41] was the first to explore the potential of KAN in vision tasks by integrating specialized KAN layers into a redesigned U-net architecture. This approach demonstrated the applicability of KAN in task-driven models. It successfully combined the structural stability of the traditional U-net architecture [42] with the innovative features of KAN, achieving substantial improvements in training and inference efficiency. Inspired by the U-KAN method, we introduced KAN layers into the text-to-image person retrieval task. Through in-depth research and optimization of the network backbone, we aimed to achieve higher accuracy while reducing computational costs, fully harnessing KAN's potential in this domain. Specifically, the integration of KAN significantly reduced the network's reliance on linear weight matrices. KAN can directly learn parameterized nonlinear activation functions, such as spline functions. This flexibility enables the network to dynamically optimize its structure based on task requirements, greatly enhancing its expressive power. Moreover, KAN improved parameter efficiency and model interpretability, effectively addressing several inherent limitations of multilayer perceptrons (MLPs). These improvements not only drive innovation in neural network architectures but also provide more efficient and stable solutions for complex tasks like text-to-image retrieval.

In practice, we integrated KAN layers into the text-image feature extraction network of VPM-Net. After the CLIP model, KAN was used to optimize text and image feature vectors. This approach significantly enhanced semantic alignment between text and image features during the feature fusion stage. At the same time, it reduced redundant computations and further improved the performance and stability of the model in text-to-image person retrieval tasks.

3.3 MINP Module

In the realm of metric learning, to mitigate the loss of ID information caused by the application of VPT (Visual Prompt Tuning) technology, the role of the ID loss function is somewhat diminished. To address this, we have meticulously designed the MINP module to focus on hard positive samples. By leveraging this module at the metric learning level, we enhance the accuracy of model adjustments. As shown in Fig. 4, the MINP module is designed to improve retrieval performance by ensuring that correct matches have lower rank values, which is crucial for a good ReID (Re-identification) system. This approach helps in reducing the proportion of incorrect samples among the retrieved results, thereby improving the overall performance of the model. In the context of cross-modal retrieval tasks, a significant challenge arises from the presence of numerous negative samples that do not match the target. These negative samples often exhibit simple structures, making them easily distinguishable, yet they provide minimal valuable information for network training. The overwhelming quantity of such easy negatives can induce a phenomenon known as "inertia", where the model becomes biased towards these samples, overshadowing a limited number of more informative positive samples that are beneficial for training [43–45]. To address this issue, we propose the **MINP** module, which integrates an original loss function with a regularization component based on a focus mechanism, specifically targeting challenging samples. The MINP module aims to enhance the model's ability to focus on difficult samples, facilitating the extraction of potentially informative yet submerged samples from the plethora of easier negatives. This process not only improves the alignment with these challenging samples but also leads to better fine-grained retrieval accuracy.

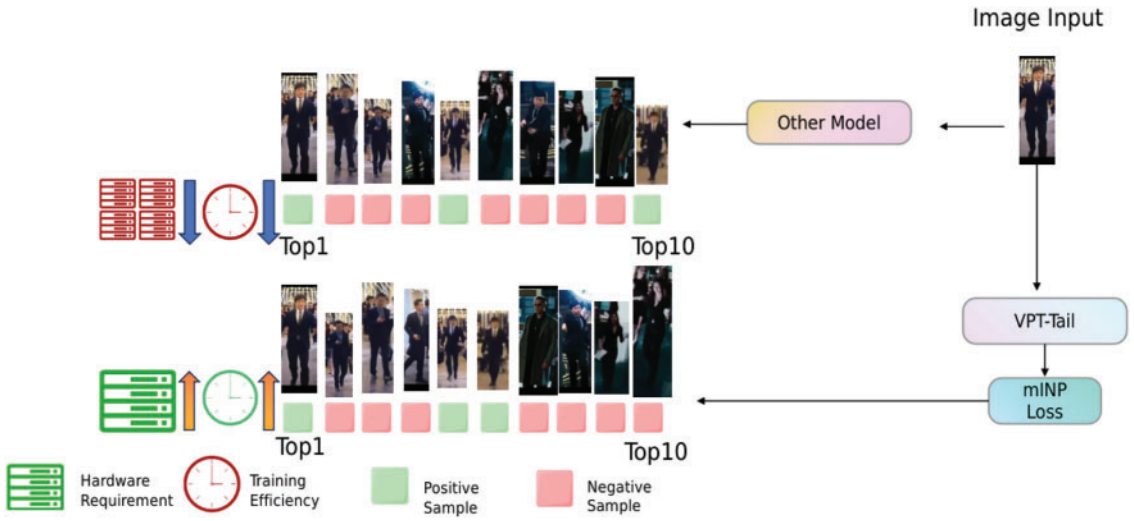

Figure 4: The operational mechanism of the MINP module. Due to the matrix operations involved in the VPT method, which introduces prompts to the model for downstream tasks, the original identity information of the samples in the model is inevitably disrupted. Therefore, as shown in the figure, after applying image prompts, we particularly focus on challenging yet valuable samples for training, to compensate for the impact of the prompt technique on identity information. Specifically, when there are many simple and easily distinguishable negative samples, the module prioritizes the more difficult parts of the positive samples, focusing from the back to the front. This ensures that these positive samples, which might otherwise be overlooked but are crucial for model training, are identified and accurately matched earlier in the process

MINP Module is an organic combination of ID loss and focal loss. Due to the application of image prompt fine-tuning techniques, the model may lose some ID information. Traditional ID loss, based on cross-entropy, computes the loss for only a single augmented image. Therefore, we introduce the focal loss function to focus on difficult positive samples in the later stages and effectively fine-tune the ID loss. As shown in Fig. 5, for difficult samples with large x values, their corresponding y values are reduced, which helps to effectively filter out those difficult positive samples that may be ignored earlier in the training process, thereby improving the model's recognition ability for these samples. Processing images with VPT usually leads to the loss of original ID information, making it unsuitable as a key attribute for local alignment. Consequently, many simple negative samples do not tightly match the target and contribute very little to network training. The abundance of such samples further complicates effective model learning, as they often overshadow more informative instances [46]. To alleviate this class imbalance, we designed the MINP loss module, inspired by the original ID loss [9] and focal loss [34]. The MINP loss includes the three components. **Focal Mechanism:** The MINP loss adopts a focal mechanism similar to focal loss, prioritizing difficult-to-match samples over simple negative samples. This is particularly beneficial for ReID tasks, where the number of unmatched samples far exceeds the matched samples. By emphasizing challenging samples, the model can better distinguish mismatched sample pairs, thereby improving retrieval accuracy. **Regularization Term:** A regularization term is added to constrain the distribution of embedding vectors. This promotes a compact embedding space, reduces the risk of overfitting, and enhances the model's generalization ability. In high-dimensional embedding spaces, this regularization ensures that the vectors remain interpretable and are not excessively dispersed. **Combined Loss Function:** By combining focal loss with the regularization term, the MINP loss effectively addresses multiple factors during the training process. Focal loss counters sample imbalance, while the regularization term helps mitigate overfitting, jointly improving the stability and robustness of the model [47].

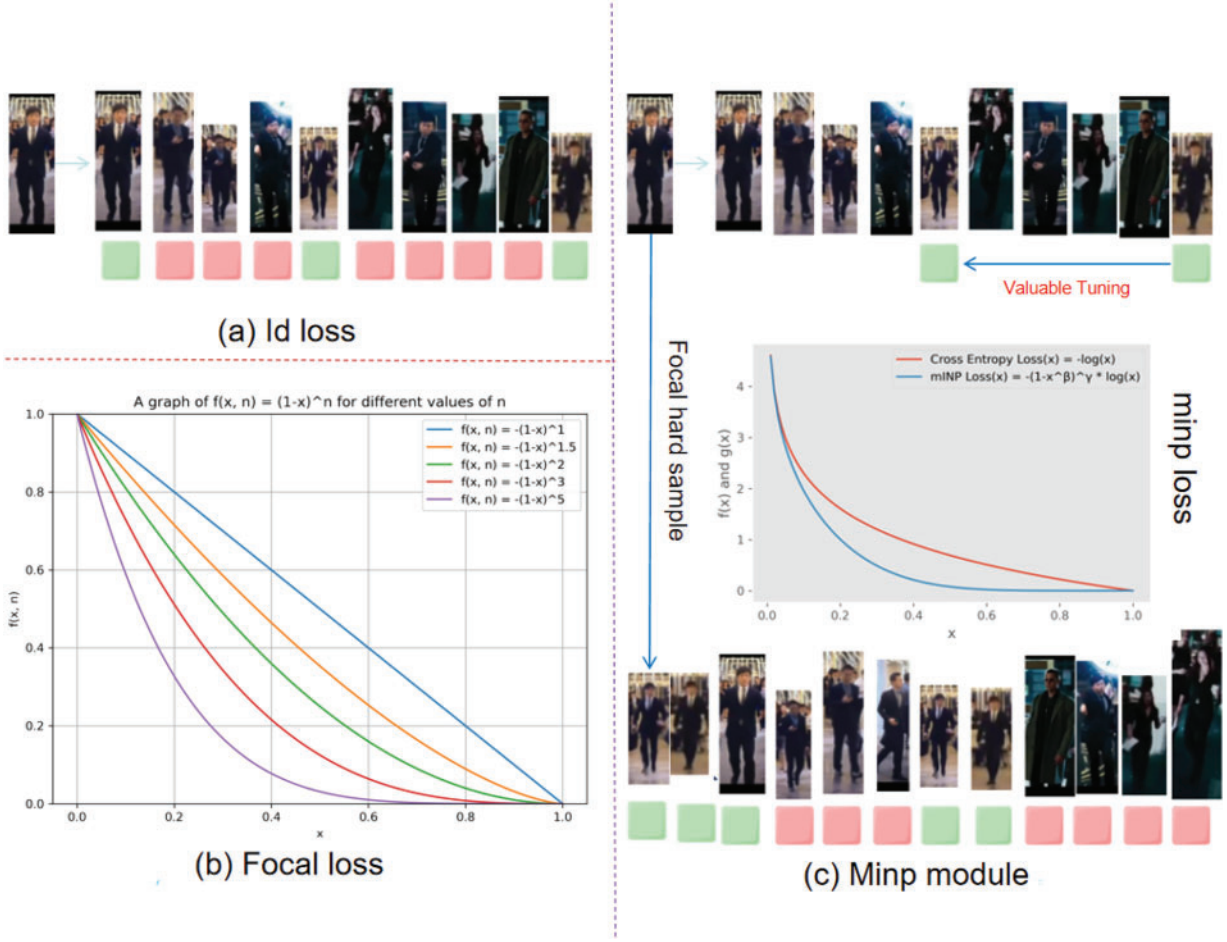

Figure 5: The component of MINP module. The MINP module is an organic combination of ID loss and focal loss. Due to the application of image prompt fine-tuning, the model may lose some identity information. Traditional ID loss, based on cross-entropy, computes the loss only for a single enhanced image. To address this, we introduce the focal loss function, which focuses on the difficult positive samples in the later stages and effectively fine-tunes the ID loss. As shown in the figure, for challenging samples with larger x-values, their corresponding y-values are reduced. This adjustment enables the model to more effectively identify difficult positive samples that might have been overlooked earlier in the training process, thereby enhancing the model's ability to recognize these samples

The implementation of the MINP Loss module involves several key steps, each contributing to enhancing the retrieval performance in the context of the ReID task. The MINP Loss incorporates both a focus mechanism and a regularization term, addressing challenges such as sample imbalance and overfitting. The steps are as follows:

Cosine Similarity Calculation: We begin by calculating the cosine similarity between the text embedding vector $\mathbf{t}_i$ and the image embedding vector $\mathbf{v}_i$ for each text-image pair i. The cosine similarity is defined as:

$$\text{CosineSimilarity}(\mathbf{t}_i, \mathbf{v}_i) = \frac{\mathbf{t}_i \cdot \mathbf{v}_i}{\|\mathbf{t}_i\| \|\mathbf{v}_i\|} \tag{12}$$

where $\mathbf{t}_i$ and $\mathbf{v}_i$ are the text and image embeddings, respectively, and $\cdot$ denotes the dot product. The cosine similarity metric, ranging from −1 to 1, quantifies the alignment between the embeddings, with higher values indicating better alignment.

Focus Loss Calculation: To prioritize difficult-to-match samples, we introduce a focus loss that leverages the similarity scores computed in the previous step. Given a set of N text-image pairs, we first calculate the similarity S_i for each pair. Then, the focus loss for each sample i is computed as follows:

$$\mathscr{L}_{\text{focus}} = -\sum_{i=1}^{N} \alpha(1 - S_i)^{\beta} \cdot y_i \cdot \log(S_i) \tag{13}$$

where S_i is the cosine similarity between the text and image embeddings for the i-th pair, y_i is the label of the i-th pair, indicating whether the pair is correctly matched ($y_i = 1$) or not ($y_i = 0$), α and β are hyperparameters controlling the focus strength and the emphasis on hard-to-match pairs. The term $(1 - S_i)^{\beta}$ penalizes pairs with low similarity, encouraging the model to focus more on difficult pairs that are harder to match.

Regularization Term Addition: A regularization term is added to the loss function to prevent overfitting and to promote the compactness of the embedding space. This term encourages smaller magnitudes for both the text and image embeddings. The regularization term is expressed as:

$$\mathscr{L}_{\text{reg}} = \lambda_1 \left(\|\mathbf{t}_i\|^2 + \|\mathbf{v}_i\|^2 \right) \tag{14}$$

where λ_1 is a regularization hyperparameter that controls the strength of the regularization term, $\|\mathbf{t}_i\|^2$ and $\|\mathbf{v}_i\|^2$ are the squared L2 norms of the text and image embedding vectors, respectively. This regularization term ensures that the embeddings do not become too large, thereby improving generalization and preventing overfitting.

Final Loss Calculation: The final loss function combines the focus loss and the regularization term. The composite loss function is given by:

$$\mathscr{L}_{\text{MINP}} = \mathscr{L}_{\text{focus}} + \mathscr{L}_{\text{reg}} \tag{15}$$

Substituting the expressions for $\mathscr{L}_{\text{focus}}$ and $\mathscr{L}_{\text{reg}}$, we obtain the final objective function to optimize:

$$\mathscr{L}_{\text{MINP}} = -\sum_{i=1}^{N} \alpha(1 - S_i)^{\beta} \cdot y_i \cdot \log(S_i) + \lambda_1 \left(\|\mathbf{t}_i\|^2 + \|\mathbf{v}_i\|^2 \right) \tag{16}$$

Thus, the MINP Loss aims to optimize the focus on challenging samples, while also promoting compact and well-distributed embeddings. By balancing these two components, the model is better able to handle imbalances in the dataset and improve its generalization ability.

4 Experiments

In this section, we present the experimental components of our study, structured into two main parts: a comparison with State-of-the-Art methods and an ablation study.

4.1 Experimental Setup

Datasets: To evaluate and validate our proposed method, we selected three challenging text-to-image retrieval datasets: CUHK-PEDES, ICFG-PEDES, and RSTPReid.

CUHK-PEDES: This dataset comprises 40,206 images and 80,412 text descriptions across 13,003 identities. Following the official dataset division, it consists of 11,003 identities, 34,054 images, and 68,108 text descriptions.The validation and test sets include 3078 and 3074 images, with corresponding text descriptions of 6158 and 6156, respectively, each containing information on 1000 identities.

ICFG-PEDES: The ICFG-PEDES dataset contains a total of 54,522 images and 4102 identities. Each image in this dataset corresponds to a single text description. After the official data split, the training set includes 34,674 image-text pairs and 3102 identities, while the test set comprises 1000 identities and 19,848 image-text pairs.

RSTPReid: RSTPReid includes 4101 identities and 20,505 images captured by 15 different cameras, with each identity associated with five images taken by different cameras. Each image is accompanied by two text descriptions. According to the official division, the dataset is split into training and test sets, containing 3701, 200, and 200 identities, respectively.

Implementation Details: VPM-Net consists of a pre-trained image encoder (CLIP-ViT-B/16), a pre-trained text encoder (CLIP text Transformer), and the VPT, KAN, and MINP modules. During training, image data augmentation methods such as random horizontal flipping, affine transformations, and random erasing are used. The size of all input images is adjusted to 384×384. The maximum length of the text token sequence L is set to 77. Our model is trained using the AdamW optimizer, with 90 training epochs and an initial learning rate of 2×10^{-5}, along with cosine learning rate decay. During the first 10 warm-up epochs, the learning rate increases linearly from 1×10^{-6} to 2×10^{-5}. For randomly initialized modules, the initial learning rate is set to 2×10^{-5}. Experiments were conducted on Unison UOS servers with the latest Ryzen 9 7995W CPU and Huawei Atlas 900 supercluster GPUs, supported by the CAAI-Huawei MindSpore Open Fund. The experiments were assisted by the FastDeploy and Development Toolkit (DTK) suites, along with dedicated drivers tailored for our experimental setup.

Evaluation Metrics: We conducted experiments on the aforementioned datasets, employing the widely recognized rank-k index (where $k = 1, 5, 10$) as our primary evaluation metric. The rank-k index quantifies the probability of retrieving at least one matching character image from the top-k candidate list given a text description. To enhance the credibility and reliability of our findings, we also utilized Mean Average Precision (mAP). Metrics-rank-k and mAP have a positive correlation with model performance, indicating that higher values reflect better model effectiveness.

4.2 Comparison with State-of-the-Art Methods

In this section, we conduct experiments on the three common benchmark datasets mentioned to compare our improved approach with State-of-the-Art methods across various dimensions, thereby illustrating the generalization capabilities of our model. To comprehensively evaluate the impact of different components in the IRRA framework, we perform an extensive empirical analysis on three public datasets: CUHK-PEDES [16], ICFG-PEDES [11], and RSTPReid [27]. The results are presented in terms of Rank-1, Rank-5, and Rank-10 accuracies (%).

Performance Comparisons on CUHK-PEDES

As shown in Table 1, our model achieves the highest Rank-1 accuracy of 72.75%, significantly outperforming other methods. For instance, CFine achieves a Rank-1 accuracy of 69.57%, while IRRA achieves 71.78%. Compared to these methods, our model surpasses CFine by 3.18% and IRRA by 0.97%. This improvement highlights the effectiveness of our approach in handling diverse and challenging image-text pairs. Additionally, our model attains competitive Rank-5 and Rank-10 results, further demonstrating its

robust generalization across different retrieval settings. These results underscore the advantages of our model in terms of both accuracy and retrieval versatility.

Table 1: Performance comparisons with state-of-the-art methods on CUHK-PEDES dataset. Results are ordered based on the Rank-1 accuracy. "G" and "L" in the "Type" column stand for the global-matching/local-matching method

Method	Type	Ref.	Image Enc.	Text Enc.	Rank-1	Rank-5	Rank-10	mAP
CMPM/C [8]	L	ECCV18	RN50	LSTM	49.37	–	79.27	–
TIMAM [48]	G	ICCV19	RN101	BERT	54.51	77.56	79.27	–
ViTAA [13]	L	ECCV20	RN50	LSTM	54.92	75.18	82.90	51.60
NAFS [49]	L	arXiv21	RN50	BERT	59.36	79.13	86.00	54.07
DSSL [27]	L	MM21	RN50	BERT	59.98	80.41	87.56	–
SSAN [11]	L	arXiv21	RN50	LSTM	61.37	80.15	86.73	–
LapsCore [26]	L	ICCV21	RN50	BERT	63.40	–	87.80	–
ISANet [30]	L	arXiv22	RN50	LSTM	63.92	82.15	87.69	–
LBUL [50]	L	MM22	RN50	BERT	64.04	82.66	87.22	–
Han et al. [33]	G	BMVC21	CLIP-RN101	CLIP-Xformer	64.08	81.73	88.19	60.08
SAF [51]	L	ICASSP22	ViT-Base	BERT	64.13	82.62	88.40	–
TIPCB [10]	L	Neuro22	RN50	BERT	64.26	83.19	89.10	–
CAIBC [25]	L	MM22	RN50	BERT	64.43	82.87	88.37	–
AXM-Net [28]	L	MM22	RN50	BERT	64.44	80.52	86.77	58.73
LGUR [29]	L	MM22	DeiT-Small	BERT	65.25	83.12	89.00	–
IVT [52]	G	ECCV22	ViT-Base	BERT	65.59	83.11	89.21	–
CFine [23]	L	arXiv22	CLIP-ViT	BERT	69.57	85.93	91.15	–
CLIP-VIT-B/16	G	CVPR23	CLIP-ViT	CLIP-Xformer	68.19	86.47	91.47	61.12
IRRA	G	CVPR23	CLIP-ViT	CLIP-Xformer	71.78	89.01	93.52	64.36
VPM-Net (Ours)	G		CLIP-VIT(prompt)	CLIP-Xformer(prompt)	72.75	89.69	93.34	64.69

Performance Comparisons on ICFG-PEDES

On the ICFG-PEDES dataset (Table 2), our approach achieves a Rank-1 accuracy of 64.04%, which outperforms IRRA (63.46%) and the baseline CLIP-VIT-B/16 (56.74%). Specifically, our method improves upon IRRA by 0.58% and outperforms CLIP-VIT-B/16 by a substantial margin of 7.3%. The improvement can be largely attributed to the integration of the MINP module, which enhances the model's ability to differentiate similar classes and reduce inter-class confusion. The mAP score further validates the effectiveness of the MINP module, as it contributes to higher retrieval precision. These results indicate that our method offers significant improvements in handling fine-grained visual-textual retrieval tasks.

Table 2: Performance comparisons with state-of-the-art methods on ICFG-PEDES dataset

Method	Type	Rank-1	Rank-5	Rank-10	mAP
Dual Path [9]	G	38.99	59.44	68.41	–
CMPM/C [8]	L	43.51	65.44	74.26	–
ViTAA [13]	L	50.98	68.79	75.78	–
SSAN [11]	L	54.23	72.63	79.53	–

(Continued)

Table 2 (continued)

Method	Type	Rank-1	Rank-5	Rank-10	mAP
IVT [52]	G	56.04	73.60	80.22	–
ISANet [30]	L	57.73	75.42	81.72	–
CFine [23]	L	60.83	76.55	82.42	–
Baseline (CLIP-ViT-B/16)	G	56.74	75.72	82.26	31.84
IRRA	G	63.46	80.25	85.82	38.06
Ours	G	64.04	80.99	85.41	38.81

Performance Comparisons on RSTPReid

In the RSTPReid dataset (Table 3), our model achieves a Rank-1 accuracy of 61.26% and an mAP score of 47.31%. Compared to CFine (50.55%) and IRRA (60.20%), our method demonstrates a significant improvement, surpassing CFine by 10.71% in Rank-1 accuracy and outperforming IRRA by 1.06%. Additionally, our model achieves an mAP score of 47.31%, surpassing CFine's mAP score of 43.41% by 3.90%. These results highlight the robustness of our approach, particularly in addressing small-object and fine-grained distinctions in the dataset. The improvements in both Rank-1 accuracy and mAP score demonstrate the effectiveness of our model in adapting to different data domains, especially in fine-grained and challenging retrieval tasks.

Table 3: Performance comparisons with state-of-the-art methods on RSTPReid dataset

Method	Type	Rank-1	Rank-5	Rank-10	mAP
DSSL [27]	G	39.05	62.60	73.95	–
SSAN [11]	L	43.50	67.80	77.15	–
LBUL [50]	L	45.55	68.20	77.85	–
IVT [52]	G	46.70	70.00	78.80	–
CFine [23]	L	50.55	72.50	81.60	–
Baseline (CLIP-VIT-B/16)	G	54.05	80.70	88.01	43.41
IRRA	G	60.20	81.30	88.20	47.17
Ours	G	61.26	80.98	88.53	47.31

In summary, the proposed model demonstrates significant advancements in cross-modal text-to-image retrieval, leveraging the innovative VPT, KAN, and MINP modules to address challenges related to diverse image-text pairs, fine-grained distinctions, and category imbalances. The performance comparisons across multiple datasets clearly illustrate the effectiveness of our approach. On the CUHK-PEDES dataset, our model achieves the highest Rank-1 accuracy of 72.75%, surpassing CFine by 3.18% and IRRA by 0.97%. This highlights the model's robustness and ability to handle complex image-text pairs. Similarly, on the ICFG-PEDES dataset, our method outperforms both IRRA and the baseline CLIP-VIT-B/16 by a significant margin, achieving a Rank-1 accuracy of 64.04% and demonstrating the power of the MINP module in improving retrieval precision through better handling of fine-grained distinctions. Finally, on the RSTPReid dataset, our approach achieves a Rank-1 accuracy of 61.26% and an mAP score of 47.31%, surpassing both CFine and IRRA in both metrics, with notable improvements in small-object and fine-grained retrieval. These results

underscore the effectiveness of the VPT, KAN, and MINP modules in enhancing the model's discriminative power and adaptability, establishing our method as a strong performer in the cross-modal retrieval domain.

5 Ablation Study

To thoroughly assess the impact of the proposed modules VPT, MINP, and KAN on the performance of our model, we conduct a detailed ablation study. The objective of this study is to evaluate how the removal or replacement of each module affects the performance across various datasets of the model. By isolating the effects of each module, this analysis not only demonstrates the contributions of the individual components but also highlights the potential benefits of combining them in the overall framework.

5.1 Single Module Analysis

The ablation study results shown in Table 4 provide a detailed comparison of the model performance under different configurations. Each module contributes to improving the model performance across all datasets.

Table 4: Ablation study on each component of VPM-Net on CUHK-PEDES, ICFG-PEDES, and RSTPReid

Methods	CUHK-PEDES			ICFG-PEDES			RSTPReid		
	Rank-1	Rank-5	Rank-10	Rank-1	Rank-5	Rank-10	Rank-1	Rank-5	Rank-10
Baseline (IRRA)	71.78	89.01	93.52	63.46	80.25	85.82	60.21	81.31	88.21
Baseline + VPT	70.11	87.64	91.93	61.55	78.48	83.19	58.77	79.02	86.93
Baseline + MINP	72.13	89.37	92.46	63.05	79.87	85.53	59.36	82.14	88.65
Baseline + KAN	71.91	89.54	92.24	62.04	79.92	85.41	60.86	80.98	87.53
Baseline + VPT + MINP	72.21	88.54	92.34	63.04	78.99	84.41	59.86	80.98	87.53
Baseline + VPT + KAN	71.86	88.24	92.07	62.93	78.65	84.11	59.29	80.17	86.39
Baseline VPT + MINP + KAN	72.75	88.69	93.74	64.04	80.99	85.41	61.26	80.98	88.53
Baseline + VPT-T	65.35	82.28	90.13	58.15	73.68	80.09	55.56	73.83	81.33
Baseline + VPT-P	70.11	87.64	91.93	61.55	78.48	83.19	58.77	79.02	86.93

VPT Module: The VPT module significantly improves training efficiency by selectively freezing the vector parameters at the tail of the model. Although this module does not directly enhance training accuracy, it plays a crucial role in improving training efficiency. This improvement arises from the module's ability to effectively preserve important learned representations in the model while allowing for flexible fine-tuning of other parts, thereby optimizing the model's ability to distinguish subtle variations in the data. As a result, the model can complete training in a shorter time while maintaining high accuracy and fully exploiting the valuable information within the training data.

The research results, as shown in Table 4, compare the performance of different VPT-Portion and VPT-Tail on three datasets for this task. VPT-Portion demonstrates significant advantages. It not only shows higher efficiency in handling the complex task of local feature alignment but also strikes an effective balance between model performance and generalization ability. The comprehensive application of the VPT-Portion technique provides a thorough and efficient solution to the challenges of cross-modal retrieval, particularly in visual-language scenarios. This indicates that the method enhances task execution efficiency while ensuring model accuracy and adaptability, making it highly competitive in practical applications.

In Table 5, an analysis from the perspectives of time cost and model adaptability reveals that although the VPT module in VPM-Net results in a slight decline in final performance, it is a crucial component for

reducing the number of parameters and improving model performance. By fine-tuning a small number of additional parameters, the module effectively adjusts the entire backbone network, significantly enhancing the model's adaptability. Therefore, the VPT module holds an irreplaceable and important position within the VPM-Net framework. From a time cost perspective, the integrated VPM-Net system demonstrates superior performance in terms of Rank-1, model performance, and overall time overhead when compared to the IRRA and benchmark frameworks. This advantage highlights that by introducing the VPT module and optimizing the overall architecture, VPM-Net achieves a good balance between performance and efficiency, making it more competitive for practical applications.

Table 5: Comparisons between different Multimodal Interaction Module of IRRA on CUHK-PEDES

Method	Param (M)	Time (ms)	Rank-1	Rank-5	Rank-10
CLIP-VIT-B	12.61	19.20	72.09	86.47	91.47
IRRA	13.66	6.42	71.78	89.01	93.52
VPT-Portion	9.06	5.99	70.11	87.64	91.93
VPM-NET	11.59	6.16	72.75	89.69	93.34

MINP Module: The introduction of the MINP module further improves the performance of the model. By focusing on hard negative samples, MINP addresses the challenge of sample imbalance in training, where the model can become overwhelmed by easy negative samples. The addition of MINP raises the Rank-1 accuracy to 72.13% on CUHK-PEDES, demonstrating its ability to identify and emphasize more challenging instances that are critical for improving the model's discriminative power.

KAN Module: The KAN (Knowledge-Aware Network) module brings external knowledge into the model, enhancing its ability to leverage supplementary information for improved decision-making. When incorporated into the model, KAN increases the Rank-1 accuracy to 71.91%. This small but significant improvement demonstrates how the integration of external knowledge can enrich the model's discrimination capabilities, particularly in distinguishing between more complex and nuanced classes in the retrieval task.

5.2 Combined Module Analysis

When the modules are combined, we observe a substantial boost in the model's performance, showcasing the synergy between the VPT, MINP, and KAN modules. Each module complements the others, and their collective impact is evident in the following results:

VPT + MINP: The combination of VPT and MINP modules results in a Rank-1 accuracy of 72.21%. This combination benefits from both the fine-tuning advantages of VPT and the focused attention on hard negative samples provided by MINP. The synergy between these two modules addresses both the model's capacity for learning effective features and its robustness in handling challenging samples, resulting in a significant performance boost.

VPT + KAN: When VPT is combined with KAN, the model achieves a Rank-1 accuracy of 71.21%. Although slightly lower than the VPT + MINP combination, this result still illustrates the benefit of integrating external knowledge with fine-tuning. The KAN module enriches the model's decision-making by bringing in external contextual knowledge, which helps the model make better predictions, even in complex scenarios.

VPT + MINP + KAN: The most significant improvement occurs when all three modules VPT, MINP, and KAN-are integrated into the model. This configuration achieves a Rank-1 accuracy of 72.75%, the highest

observed in our study. The combination of VPT's fine-tuning, MINP's handling of hard negatives, and KAN's external knowledge create a robust model that excels at distinguishing between complex classes while maintaining high efficiency in handling imbalanced data. The collective effect of these modules highlights their complementary nature, where each module addresses a specific challenge in the person re-identification and cross-modal retrieval tasks.

Our ablation study demonstrates that each module VPT, MINP, and KAN-provides a meaningful improvement to the model's performance. The VPT module enhances fine-tuning capabilities, allowing the model to better adapt to task-specific data. MINP addresses the challenge of sample imbalance by focusing on harder negatives, thus preventing the model from being dominated by easy negatives during training. Meanwhile, the KAN module leverages external knowledge to improve the model's discriminative power, enriching its ability to distinguish complex instances.

6 Conclusion

In this paper, we address the issues of complex model structures and high adjustment costs in the field of cross-modal retrieval by ingeniously combining VPT (Visual Prompt Tuning) technology with the CLIP pre-trained model for cross-modal person re-identification. This approach not only enhances model performance by pre-mapping images and text into a unified joint space for holistic processing but also achieves efficient model adjustment through the fine-tuning of a small number of additional parameters, significantly improving the model's adaptability and reducing adjustment costs. In terms of methodology, we have innovated upon existing VPT techniques by designing VPT-Tail and VPT-Portion. Notably, the VPT-Portion method strikes an optimal balance between model performance and computational efficiency. Furthermore, recognizing that the enhancement of model performance and adaptability through VPT technology could potentially compromise model stability and precision, we have reinforced the stability and accuracy of model adjustments from two dimensions: representation learning and metric learning. Specifically, the Knowledge-Aware Network (KAN) module unifies the representation of linear and non-linear relationships in representation learning, while the Multi-Instance Negative Pooling (MINP) module enhances the model's ability to learn from hard positive samples in metric learning. The resulting VPM-Net framework is more stable and accurate. Experimental results demonstrate that each component of the VPM-NET is indispensable and that the framework outperforms existing methods across various datasets. Looking ahead, we plan to expand the application of our VPM-NET framework to larger and more diverse datasets to evaluate the generalizability of the proposed methods.

Acknowledgement: The authors are grateful to the HERO Laboratory, Hubei University of Technology, Wuhan, China for support.

Funding Statement: This research was funded by the Key Research and Development Program of Hubei Province, China (Grant No. 2023BEB024), the Young and Middle-aged Scientific and Technological Innova-tion Team Plan in Higher Education Institutions in Hubei Province, China (Grant No. T2023007) and the key projects of Hubei Provincial Department of Education (No. D20161403).

Author Contributions: Conceptualization: Haitao Xie, Yuliang Chen, Yunjie Zeng; Methodology: Haitao Xie, Lingyu Yan, Zhizhi Wang; Software: Haitao Xie, Yuliang Chen, Yunjie Zeng; Validation: Haitao Xie, Yuliang Chen, Yunjie Zeng; Writing—original draft preparation: Haitao Xie, Yuliang Chen, Zhizhi Wang; Writing—review and editing: Haitao Xie, Lingyu Yan, Zhizhi Wang; Visualization: Lingyu Yan, Zhizhi Wang, Zhiwei Ye; Supervision: Haitao Xie, Yunjie Zeng; Project administration: Lingyu Yan, Yunjie Zeng, Zhizhi Wang. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The datasets generated and/or analyzed during the current study are available from the corresponding author on reasonable request.

Ethics Approval: Not applicable.

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

References

1. He S, Luo H, Wang P, Wang F, Li H, Jiang W. TransReID: transformer-based object re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2021. p. 15013–22.
2. Luo H, Gu Y, Liao X, Lai S, Jiang W. Bag of tricks and a strong baseline for deep person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops; 2019.
3. Wang H, Shen J, Liu Y, Gao Y, Gavves E. NFormer: robust person re-identification with neighbor transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2022. p. 7297–307.
4. Farooq A, Awais M, Kittler J, Akbari A, Khalid SS. Cross modal person re-identification with visual-textual queries. In: 2020 IEEE International Joint Conference on Biometrics (IJCB); 2020; IEEE. p. 1–8.
5. Lin D, Peng Y, Meng J, Zheng WS. Cross-modal adaptive dual association for text-to-image person retrieval. IEEE Trans Multimed. 2024;26:6609–20.
6. Bao L, Wei L, Qiu X, Zhou W, Li H, Tian Q. Learning transferable pedestrian representation from multimodal information supervision. arXiv:2304.05554. 2023.
7. Chen Y, Huang R, Chang H, Tan C, Xue T, Ma B. Cross-modal knowledge adaptation for language-based person search. IEEE Trans Image Process. 2021;30:4057–69. doi:10.1109/TIP.2021.3068825.
8. Zhang Y, Lu H. Deep cross-modal projection learning for image-text matching. In: Proceedings of the European Conference on Computer Vision (ECCV); 2018. p. 686–701.
9. Zheng Z, Zheng L, Garrett M, Yang Y, Xu M, Shen YD. Dual-path convolutional image-text embeddings with instance loss. ACM Trans Multimed Comput Commun Appl. 2020;16(2):1–23. doi:10.1145/3383184.
10. Chen Y, Zhang G, Lu Y, Wang Z, Zheng Y. TIPCB: a simple but effective part-based convolutional baseline for text-based person search. Neurocomputing. 2022;494(2):171–81. doi:10.1016/j.neucom.2022.04.081.
11. Ding Z, Ding C, Shao Z, Tao D. Semantically self-aligned network for text-to-image part-aware person re-identification. arXiv:2107.12666. 2021.
12. Jing Y, Si C, Wang J, Wang W, Wang L, Tan T. Pose-guided multi-granularity attention network for text-based person search. Proc AAAI Conf Artif Intell. 2020;34(7):11189–96. doi:10.1609/aaai.v34i07.6777.
13. Wang Z, Fang Z, Wang J, Yang Y. Vitaa: visual-textual attributes alignment in person search by natural language. In: European Conference on Computer Vision; 2020; Springer. p. 402–20.
14. Jia M, Tang L, Chen BC, Cardie C, Belongie S, Hariharan B, et al. Visual prompt tuning. In: European Conference on Computer Vision; 2022; Springer. p. 709–27.
15. Cao M, Bai Y, Zeng Z, Ye M, Zhang M. An empirical study of clip for text-based person search. Proc AAAI Conf Artif Intell. 2024;38(1):465–73. doi:10.1609/aaai.v38i1.27801.
16. Li S, Xiao T, Li H, Zhou B, Yue D, Wang X. Person search with natural language description. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2017. p. 1970–9.
17. Jiang D, Ye M. Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2023. p. 2787–97.
18. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556. 2014.
19. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput. 1997;9(8):1735–80. doi:10.1162/neco.1997.9.8.1735.
20. Chen T, Xu C, Luo J. Improving text-based person search by spatial matching and adaptive threshold. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV); 2018; IEEE. p. 1879–87.

21. Li S, Xiao T, Li H, Yang W, Wang X. Identity-aware textual-visual matching with latent co-attention. In: Proceedings of the IEEE International Conference on Computer Vision; 2017. p. 1890–9.
22. Chen YC, Li L, Yu L, El Kholy A, Ahmed F, Gan Z, et al. UNITER: Universal image-text representation learning. In: European Conference on computer vision; 2020; Springer. p. 104–20.
23. Yan S, Dong N, Zhang L, Tang J. CLIP-driven fine-grained text-image person re-identification. arXiv:2210.10276. 2022.
24. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805. 2018.
25. Wang Z, Zhu A, Xue J, Wan X, Liu C, Wang T, et al. CAIBC: capturing all-round information beyond color for text-based person retrieval. arXiv:2209.05773. 2022.
26. Wu Y, Yan Z, Han X, Li G, Zou C, Cui S. Language-guided person search via color reasoning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2021. p. 1624–33.
27. Zhu A, Wang Z, Li Y, Wan X, Jin J, Wang T, et al. DSSL: deep surroundings-person separation learning for text-based person retrieval. In: Proceedings of the 29th ACM International Conference on Multimedia; 2021. p. 209–17.
28. Farooq A, Awais M, Kittler J, Khalid SS. AXM-Net: implicit cross-modal feature alignment for person re-identification. Proc AAAI Conf Artif Intell. 2022;36(4):4477–85.
29. Shao Z, Zhang X, Fang M, Lin Z, Wang J, Ding C. Learning granularity-unified representations for text-to-image person re-identification. arXiv:2207.07802. 2022.
30. Yan S, Tang H, Zhang L, Tang J. Image-specific information suppression and implicit local alignment for text-based person search. arXiv:2208.14365. 2022.
31. Wang C, Ning X, Li W, Bai X, Gao X. 3D person re-identification based on global semantic guidance and local feature aggregation. IEEE Trans Circuits Syst Video Technol. 2023;32(5):3164–77.
32. Yu Z, Li L, Xie J, Wang C, Li W, Ning X. Pedestrian 3D shape understanding for person re-identification via multi-view learning. IEEE Trans Circuits Syst Video Technol. 2024;34(7):5589–602. doi:10.1109/TCSVT.2024.3358850.
33. Han X, He S, Zhang L, Xiang T. Text-based person search with limited data. arXiv:211010807. 2021.
34. Lin TY, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision; 2017. p. 2980–8.
35. Li X, Yin X, Li C, Zhang P, Hu X, Zhang L, et al. Oscar: object-semantics aligned pre-training for vision-language tasks. 2020. doi:10.48550/arXiv.2004.06165.
36. Desai K, Johnson J. VirTex: learning visual representations from textual annotations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2021. p. 11162–73.
37. Miech A, Alayrac JB, Laptev I, Sivic J, Zisserman A. Thinking fast and slow: efficient text-to-visual retrieval with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2021. p. 9826–36.
38. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning; 2021; PMLR. p. 8748–63.
39. Sennrich R, Haddow B, Birch A. Neural machine translation of rare words with subword units. arXiv:1508.07909. 2015.
40. Liu Z, Wang Y, Vaidya S, Ruehle F, Halverson J, Soljacic M, et al. KAN: kolmogorov-arnold networks. arXiv:2404.19756. 2024.
41. Li C, Liu X, Li W, Wang C, Liu H, Yuan Y. U-KAN makes strong backbone for medical image segmentation and generation. arXiv:2406.02918. 2024.
42. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention-MICCAI 2015: 18th International Conference; 2015 Oct 5–9; Munich, Germany: Springer; 2015. p. 234–41.
43. Sohn K. Improved deep metric learning with multi-class n-pair loss objective. In: Advances in Neural Information Processing Systems 29; 2016. p. 1849–57.
44. Wu CY, Manmatha R, Smola AJ, Krahenbuhl P. Sampling matters in deep embedding learning. In: Proceedings of the IEEE International Conference on Computer Vision; 2017. p. 2840–8.

45. Oh Song H, Jegelka S, Rathod V, Murphy K. Deep metric learning via facility location. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2017. p. 5382–90.
46. Ye M, Shen J, Lin G, Xiang T, Shao L, Hoi SC. Deep learning for person re-identification: a survey and outlook. IEEE Trans Pattern Anal Mach Intell. 2021;44(6):2872–93. doi:10.1109/TPAMI.2021.3054775.
47. Rostami AM, Homayounpour MM, Nickabadi A. Efficient attention branch network with combined loss function for automatic speaker verification spoof detection. Circuits Syst Signal Process. 2023;42(7):4252–70. doi:10.1007/s00034-023-02314-5.
48. Sarafianos N, Xu X, Kakadiaris IA. Adversarial representation learning for text-to-image matching. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2019. p. 5814–24.
49. Gao C, Cai G, Jiang X, Zheng F, Zhang J, Gong Y, et al. Contextual non-local alignment over full-scale representation for text-based person search. arXiv:2101.03036. 2021.
50. Wang Z, Zhu A, Xue J, Wan X, Liu C, Wang T, et al. Look before you leap: improving text-based person retrieval by learning a consistent cross-modal common manifold. In: Proceedings of the 30th ACM International Conference on Multimedia; 2022. p. 1984–92.
51. Li S, Cao M, Zhang M. Learning semantic-aligned feature representation for text-based person search. In: ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2022; IEEE. p. 2724–8.
52. Shu X, Wen W, Wu H, Chen K, Song Y, Qiao R, et al. See finer, see more: implicit modality alignment for text-based person retrieval. arXiv:2208.08608. 2022.