CLIP-Driven Fine-Grained Text-Image Person Re-Identification

Shuanglin Yan, Neng Dong, Liyan Zhang, and Jinhui Tang, Senior Member, IEEE

Abstract—Text-Image Person Re-identification (TIReID) aims to retrieve the image corresponding to the given text query from a pool of candidate images. Existing methods employ prior knowledge from single-modality pre-training to facilitate learning, but lack multi-modal correspondence information. Vision-Language Pre-training, such as CLIP (Contrastive Language-Image Pre-training), can address the limitation. However, CLIP falls short in capturing fine-grained information, thereby not fully leveraging its powerful capacity in TIReID. Besides, the popular explicit local matching paradigm for mining fine-grained information heavily relies on the quality of local parts and cross-modal interaction/guidance, leading to intra-modal information distortion and ambiguity problems. Accordingly, in this paper, we propose a CLIP-driven Fine-grained Information excavaTion framework (CFine) to fully utilize the powerful knowledge of CLIP for TIReID. To transfer the multi-modal knowledge effectively, we conduct fine-grained information excavaTion to mine modality-shared discriminative details for global alignment. Specifically, we propose a multi-level global feature learning (MGF) module that fully mines the discriminative local information within each modality, thereby emphasizing identity-related discriminative clues through enhanced interaction between global image (text) and informative local patches (words). MGF generates a set of enhanced global features for later inference. Furthermore, we design cross-grained feature refinement (CFR) and fine-grained correspondence discovery (FCD) modules to establish cross-modal correspondence at both coarse and fine-grained levels (image-word, sentence-patch, word-patch), ensuring the reliability of informative local patches/words. CFR and FCD are removed during inference to optimize computational efficiency. Extensive experiments on multiple benchmarks demonstrate the superior performance of our method in TIReID.

Index Terms—Text-image person re-identification, multi-modal correspondence information, intra-modal information distortion, fine-grained information excavaTion.

I. INTRODUCTION

PERSON Re-identification (ReID) is a popular and challenging task in computer vision. Over the past decade, ReID has made remarkable progress [1], [2], [3], [4], [5], and has been successfully applied in some practical scenarios. However, most existing ReID approaches assume that pedestrian images can be captured across disjoint cameras, thereby overlooking situations where pedestrian images are unavailable due to complex or special scenes, such as remote roads without cameras or instances where pedestrians are completely occluded. Although pedestrian images are not available, we can find some witnesses at the scene and search for the target pedestrian by the witness’s language description, that is, text-image person re-identification (TIReID) [6]. TIReID holds immense practical value, leading to a growing interest from both academia and industry.

TIReID is a fine-grained cross-modal retrieval task that relies on effectively mining detailed information from images and texts while establishing their correspondences. In recent years, many effective methods [7], [8], [9], [10], [11], [12] have been proposed. These methods typically leverage external knowledge to facilitate learning, often initializing the image/text backbone through single-modality pre-training (e.g., the pre-trained ResNet [13] and ViT [14] on ImageNet, the pre-training language model BERT [15]), which lacks multi-modal correspondence information. Recently, visual-language pre-training (VLP) has achieved remarkable success in learning semantically rich visual concepts with natural language supervision, and the most representative work is Contrastive Language-Image Pre-training (CLIP) [16]. To overcome the limitations of single-modality pre-training, we attempt to benefit from the abundant multi-modal knowledge of CLIP for TIReID. Fine-tuning CLIP directly on TIReID is effective. However, as depicted in Figure 1, CLIP is primarily trained to focus on instance-level representation (image-level, sentence-level), which limits its ability to capture and match fine-grained information. While TIReID requires the model to focus on intra-modal fine-grained information and inter-modal fine-grained correspondences to distinguish the subtle differences between pedestrians. The task gap, which lacks sensitivity to fine-grained information, prevents us from fully exploiting the powerful capabilities of CLIP in image-text matching. Thus, it becomes crucial to address this task gap by mining and matching fine-grained information.

To capture fine-grained matching across modalities, most existing methods adopt the explicit local matching paradigm, which is based on explicitly acquired local parts for local matching through cross-modal inter-part interaction [9], [17], [18] or guidance [11], [12]. The effectiveness of these methods is closely tied to the quality of the local parts. However, as depicted in Figure 1, the image parts corresponding to textual entities are irregular. Unfortunately, existing methods
To this end, we propose a novel CLIP-driven fine-grained information excavation framework, namely CFine, for TIReID. As illustrated in Figure 1, CFine consists of two parts: modality-specific feature extraction and fine-grained information excavation. The modality-specific encoders, initialized by CLIP, extract the image and text representations. The fine-grained information excavation part is then utilized to mine the modality-shared discriminative details within each modality, enhancing the global features for global alignment. Unlike previous methods that heavily rely on the quality of local parts and cross-modal inter-part interaction/guidance, our approach aims to enhance global alignment by incorporating fine-grained discriminative information into global features, without acquiring parts and complex interaction/guidance. Specifically, we propose a multi-level global feature learning (MGF) module to effectively capture the identity-related subtle clues within each modality. In this module, we devise a token selection process that picks out a set of informative tokens (discriminative patches/words) based on the self-attention score between the class token and local tokens for each modality. These informative tokens are then divided into multiple subsets and inputted into a global-local decoder (GLD) to generate a set of multi-level global features by enhancing the interaction between global image (text) and local discriminative patches (words). In the above process, the reliability of the selected informative tokens is crucial. However, due to the discriminative limitations of the class token, it is possible for some non-modality-shared tokens to be inadvertently selected, thereby emphasizing information that is not conducive to text-image matching.

To tackle this issue, we design two modules: the cross-grained feature refinement (CFR) module and the fine-grained correspondence discovery (FCD) module. These modules are designed to emphasize the inter-modal correspondence between selected informative tokens from different modalities through cross-modal interaction, ensuring the reliability and modality-sharing of these tokens. The CFR module filters out non-modality-shared informative tokens guided by the coarse-grained global token. It achieves this by computing cross-grained similarities (image-word, sentence-patch) between modalities, establishing the rough cross-modal correspondence. And the FCD module focuses on discovering the relationship between selected informative tokens from different modalities to establish inter-modal fine-grained correspondence. Note that the CFR and FCD modules perform parameter-free cross-modal interaction, serving as two regularization terms to constrain the behavior of the MGF module. During inference, CFR and FCD are removed, and multi-level global image and text features generated by MGF are used for cross-modal retrieval. Our main contributions are summarized as follows:

- We propose a CLIP-driven fine-grain information excavation framework to take full advantage of rich multi-modal knowledge from CLIP for TIReID.
- We achieve fine-grained text-image alignment via three innovative modules, i.e., multi-level global feature learning, cross-grained feature refinement, and fine-grained correspondence discovery, without local parts acquisition and complex interaction/guidance.
- We conduct extensive experiments on three benchmarks to validate the effectiveness of our CFine. CFine performs significantly better than previous methods and outperforms the previous SOTA method by +5.13%, +4.79%, and +3.85% on the CUHK-PEDES, ICFG-PEDES, and RSTReid, respectively.

The remainder of the paper is organized as follows. We first review the related works in Section II; Section III describes the proposed CFine in detail; Section IV reports extensive experimental results and analysis; and finally the paper is summarized in Section V.

II. RELATED WORK

A. Text-Image Person Re-Identification

TIReID, a fine-grained cross-modal retrieval task initially proposed by [6], presents unique challenges compared to
general cross-modal retrieval tasks due to its fine-grained nature. The key to TIREID is cross-modal alignment, and existing methods can be broadly classified into two classes based on their alignment strategy: cross-modal interaction-based and cross-modal interaction-free methods. Cross-modal interaction-based methods [6], [9], [10], [17], [18], [20], [21], [22] focus on mining local correspondences between image and text parts by using attention mechanism to predict the matching score for image-text pairs. Cross-modal interaction is a double-edged sword, and its advantage lies in better aligning image-text pairs and reducing the modality gap through sufficient cross-modal inter-part pairwise interaction, acquiring superior performance. But its disadvantage is the high computational cost, which greatly reduces the practicability of such methods.

For cross-modal interaction-free methods [11], [12], [23], [24], [25], [26], [27], [28], [29], [30], [31], [32], several works [23], [24], [25], [26], [27] mainly focused on designing network and optimization loss to learn globally aligned image and text embeddings in a joint latent space. These methods are efficient but their performance is not satisfactory due to the lack of fine-grained alignment. In recent years, some effective and lightweight models [11], [12], [28] have been proposed to generate locally aligned text parts guided by image parts, which achieve superior performance without cross-modal inter-part pairwise interaction. With the success of Transformer in various visual and language tasks, several Transformer-based methods [30], [31], [32] have been proposed recently and achieved state-of-the-art performance. All in all, TIREID has achieved remarkable progress over the past few years.

However, most existing methods model fine-grained matching between modalities based on the explicit local matching paradigm, which heavily relies on the quality of local parts and cross-modal inter-part interaction/guidance, resulting in intra-modality information distortion and ambiguity issues. Unlike them, we focus on mining the modality-shared discriminative details within each modality to enhance global features for global alignment without acquiring parts and complex interaction/guidance. Besides, existing TIREID methods employ prior knowledge from single-modality pre-training to facilitate learning, lacking multi-modal correspondence information. The vision-language pre-training models like CLIP [16] can address the limitation. Thus, we explore leveraging the powerful knowledge of the CLIP model for TIREID in this paper. Note that TextReID [27] is the first work using CLIP on TIREID. But they only focus on exploring the network initialization strategy of CLIP to address the problem of limited training data, without digging into the powerful capabilities of CLIP. In contrast, we delve into a comprehensive exploration of how to leverage the rich knowledge within CLIP, enabling the TIREID task to fully benefit from its capabilities.

B. Vision-Language Pre-Training

The paradigm of “pre-training and fine-tuning” has emerged as one of the most important paradigms in computer vision community. This paradigm involves initializing a model with pre-trained model parameters on large-scale datasets, followed by fine-tuning it on various downstream tasks. In this paradigm, the quality of the pre-training model plays a vital role in the optimization difficulty and performance of the downstream tasks. In the past decade, the pre-training model in single modal domain [13], [14], [15] has achieved great success. Recently, many works [16], [33], [34], [35], [36], [37] have sought to extend the concept of pre-training to the multi-modal domain, known as visual-language pre-training (VLP), leading to significant progress. These VLP models can be divided into two categories according to the pre-training tasks: (1) image-text contrastive learning tasks, which align images and texts into a shared space through cross-modal contrastive loss, e.g., CLIP [16], ALIGN [33], FILIP [34]; (2) language modeling based task, some auxiliary tasks (Masked Language/Region Modeling, image captioning, text-grounding image generation) are used to establish the correspondences between images and texts, e.g., VisualBERT [36], UNITER [37]. The adoption of VLP models has demonstrated significant performance gains in cross-modal tasks and fine-grained visual tasks, as evidenced by numerous subsequent works. Building upon this success, we anticipate leveraging the abundant multi-modal knowledge of VLP to further advance the field of TIREID.

C. CLIP-Based Fine-Tuning

As one of the most prominent models in the field of VLP, Contrastive Language-Image Pre-training (CLIP) [16] has garnered significant attention. Unlike traditional image-based supervised pre-training models, CLIP employs natural language as a supervisory signal for learning visual features through contrastive learning on web-scale image-text data. Benefiting from semantic-level language supervision, the visual network can learn high-quality visual features with rich semantic information, which has an impressive positive impact on cross-modal tasks and fine-grained visual tasks. Recently, a lot of follow-ups [38], [39], [40], [41], [42], [43], [44], [45] have been put forward to fine-tune CLIP to various downstream tasks. The most common is to adapt CLIP to some cross-modal tasks, such as video-text retrieval [38], [39], [40], video caption [41], referring image segmentation [42]. Besides, some efforts have recognized the semantic-level and high-quality visual concept representation capacity of CLIP and applied it to some fine-grained visual tasks, including dense prediction [45], point cloud understanding [44], video recognition [43], and achieved impressive results. As a cross-modal retrieval task [46] as well as a fine-grained recognition task [47], [48], [49], TIREID can also benefit from CLIP. Thus, in this paper, we try to explore an effective framework to fully transfer the CLIP model to the TIREID task.

III. METHODS

A. Overview of CFine

In this section, we elaborate on the implementation details of the proposed CFine. An overview of CFine is illustrated in Figure 2. Given a set of pedestrian images \( V \) and text descriptions \( T \), we first feed them to the dual encoders
initialized by CLIP to extract the image and text features. Second, to better adapt to TIReID, the fine-grained information excavation is performed to mine the modality-shared discriminative details within each modality to enhance global features for global alignment. For fine-grained information excavation, three modules are proposed: (1) The multi-level global feature learning (MGF) module mines discriminative local clues within each modality based on informative tokens at different levels. It generates a set of multi-level discriminative global features. (2) The cross-grained feature refinement (CFR) module is designed to filter out unnecessary information in selected informative tokens, ensuring the reliability of informative tokens. (3) The fine-grained correspondence discovery (FCD) module establishes fine-grained correspondences between informative tokens from images and texts. Finally, we compute the similarity between the learned image and text representations using the cosine similarity function, aiming to maximize the similarity if the image and text are matched and minimize it otherwise.

B. Dual Encoders

The architecture of CLIP is illustrated in Figure 1 and consists of an image encoder and a text encoder. Each encoder comprises a feature extractor and a projector. The image and text feature extractors extract features through a ViT with a width of 768 and a Transformer with a width of 512 respectively, while the projectors map image and text features to a 512-dimensional latent space, aligning the image and text representations using a contrastive objective. Fine-tuning CLIP directly on the TIReID dataset is one approach, but several works [50], [51] have shown that the projectors may lead to intra-modal information distortion. This is undesirable for TIReID which relies on fine-grained information, especially images. If the projector is removed, the dimensions of the two cannot be unified. Thus, in the paper, we solely utilize the image encoder of CLIP with the projector removed as our image encoder. For text, we use another pre-training language model BERT [15] as the text encoder. In addition, to ensure a fair comparison with existing methods, we also use ViT pre-trained on ImageNet [52] as the image encoder.

1) Image Representation: Given an image $I \in \mathcal{V}$, a visual tokenization process is first performed to convert the image to a discrete token sequence of length $N_v$. A learnable $[CLS^v]$ token is attached to the beginning of the sequence as an image-level representation. Finally, the token sequence of length $N_v + 1$ is fed into the transformer of ViT. The output of
image encoder is represented as \( V = \{v_g, v_1, v_2, \ldots, v_{N_g}\} \in \mathbb{R}^{(N_g+1) \times d} \), where \( v_g \) is the image-level global feature, \( \{v_1, v_2, \ldots, v_{N_g}\} \) is the patch-level local features.

2) Text Representation: For a text \( T \in T \), we employ the pre-trained BERT [15] as the text encoder to generate the text representation. The text \( T \) is initially tokenized using lower-cased byte pair encoding (BPE) with a 30522 vocabulary size. Then, the tokenized sequence is padded with \([CLS]\) token at the beginning. Finally, the token sequence of length \( N_I + 1 \) is fed into the text encoder to generate the sentence-level global feature \( t_g \in \mathbb{R}^d \) and the word-level local features \( \{t_1, t_2, \ldots, t_{N_I}\} \in \mathbb{R}_{N_I \times d} \).

Considering the substantial gap between the upstream pre-training task and the downstream TIReID task, \( v_g \) and \( t_g \) solely capture instance-level modality information, lacking fine-grained details that are essential for TIReID. Therefore, in the following, we conduct fine-grained information excavation (MGF, CFR, and FCD) to fully mine the modality-shared discriminative fine-grained clues within each modality.

C. Multi-Level Global Feature Learning

Given the subtle visual variations among different pedestrians in ReID, it is crucial to fully mine the fine-grained information of images/texts for distinguishing different pedestrians. Most existing TIReID methods mine fine-grained information based on explicitly acquired local parts through cross-modal inter-part interaction/guidance. These methods are not only dominated by unqualified local parts, leading to ambiguity and introducing noise, but also require high storage and computation resources. Unlike them, our method aims to mine discriminative fine-grained information at different levels and integrate it into global features, enabling the learning of discriminative global features at multiple levels without the need for local parts acquisition and complex inter-part interaction/guidance. Benefiting from the global dependency modeling capability of self-attention, Transformer achieves impressive results on various tasks. However, self-attention treats each local token equally in calculating attention weights and generates a global feature by computing a weighted sum of all local tokens. The global feature is dominated by all local tokens, and this way of considering all local tokens simultaneously reduces the influence of critical local tokens. Particularly in fine-grained recognition tasks, the way can result in a serious loss of discriminative information. To address this limitation, we propose a solution that selects informative tokens to learn global features, instead of considering all tokens simultaneously. These informative tokens are passed through the global-local decoder to learn a set of multi-level global features. By focusing on informative tokens, we enhance the discriminative power of global features and alleviate the loss of critical fine-grained information.

1) Token Selection: The class token as the output of Transformer is used for classification or retrieval, which is obtained by weighted aggregation of all local tokens. The weights reflect the correlation between the class token and each local token, with larger weights indicating greater contributions and importance to the task at hand. Therefore, we select informative tokens based on the correlation between local tokens and the class token [53], [54], [55].

Specifically, the self-attention of each Transformer block can generate an attention map of size \((1 + N) \times (1 + N)\), reflecting the correlation among the input \(1 + N\) tokens (the first is the class token). The first row of the attention map represents the dependency between the class token and local tokens. In this paper, we use the attention map \( A \in \mathbb{R}^{(1+N) \times (1+N)} \) generated by the self-attention of the last Transformer block, and the correlation score between the class token and local tokens is \( m = A[0, 1:] \in \mathbb{R}^N \). To construct a new discriminative local token sequence, we select the top 2K tokens from the \( N \) local tokens output by Transformer, corresponding to the top 2K highest scores in \( m \). The token selection process is shown in Figure 3(a).

We perform the token selection process separately for both images and texts. For image \( I \), 2K\( _i \) tokens are chosen from \( \{v_1, v_2, \ldots, v_{N_i}\} \). For text \( T \), 2K\( _t \) tokens are chosen from \( \{t_1, t_2, \ldots, t_{N_t}\} \). The selected image and text token sequences are denoted as \( V^s = \{v_1^s, v_2^s, \ldots, v_{K^i}^s\} \in \mathbb{R}^{2K_i \times d} \) and \( T^s = \{t_1^s, t_2^s, \ldots, t_{K^t}^s\} \in \mathbb{R}^{2K_t \times d} \), where \( H_{s} \) are then fed into GLD. The attention map represents the dependency between the class token \( V \) and \( m = A[0, 1:] \in \mathbb{R}^N \). To construct a new discriminative local token sequence, we select the top 2K highest scores in \( m \). The token selection process is shown in Figure 3(a).

2) Global-Local Decoder: We design a global-local decoder (GLD) to enhance the discriminative local information within images and texts, thereby improving the discriminability of global features. As depicted in Figure 3(b), the GLD consists of \( M \) blocks and generates a set of multi-level discriminative global features using the selected token sequence as input. The selected token sequence only contains informative tokens, discarding redundant ones. To fully mine discriminative fine-grained information, we divide the selected token sequence into two sub-sequences, which correspond to different discriminative levels. The former is a high-level discriminative sequence with the top \( K \) most informative tokens, while the latter is a middle-level discriminative sequence with the remaining \( K \) informative tokens. These two sub-sequences are fed into the GLD to highlight local information at different levels. Specifically, taking image as an example, \( V^s \in \mathbb{R}^{2K_i \times d} \) is divided into the high-level sequence \( V^s_h \in \mathbb{R}^{K_i \times d} \) and middle-level sequence \( V^s_m \in \mathbb{R}^{K_i \times d} \), which are respectively preprocessed with a \([CLS]^h\) token and a \([CLS]^m\) token, denoted \( v_{CLS}^s_h \in \mathbb{R}^d \) and \( v_{CLS}^s_m \in \mathbb{R}^d \), are then fed into GLD. In one GLD block, \( V^s_h/V^s_m \) is initially passed through the multi-head self-attention (MHSA) layer to propagate the information of these informative tokens into the class token.

\[
\hat{V}_h^s/\hat{V}_m^s = \text{MHSA}(\text{Norm}(\hat{V}_h^s/\tilde{V}_m^s)) + \hat{V}_h^s/\tilde{V}_m^s \quad (1)
\]

where \( \hat{V}_h^s/\tilde{V}_m^s = [v_{CLS}^s_h, V^s_h]/[v_{CLS}^s_m, V^s_m] \). \text{Norm}() denotes Layer Normalization. After that, \( \hat{V}_h^s \) and \( \hat{V}_m^s \) are fed into the multi-head cross-attention (MHCA) layer to compute the cross-attention between \( \hat{V}_h^s/V^s_m \) and \( V \), which can highlight not only the informative tokens themselves but also other associated contextual information. Finally, the output of MHCA is further fed into the multi-layer perceptron (MLP) layer to generate multiple discriminative global features at
different levels.

\[
\begin{align*}
\hat{V}_h/V_m &= M H C A ( N o r m ( \hat{V}_h/V_m ), V ) + \hat{V}_h/V_m \\
V_h/V_m &= M L P ( N o r m ( \hat{V}_h/V_m ) ) + \hat{V}_h/V_m
\end{align*}
\]  

where \( V_h/V_m \) denotes the output of a GLD block for input \( V_h/V_m \). The \([CLS]_h^V\) and \([CLS]_m^V\) tokens output by the last block are used as the high-level global image feature \( v_h \in \mathbb{R}^d \) and middle-level global image feature \( v_m \in \mathbb{R}^d \) respectively. Besides, the image-level feature \( v_g \) that treats all local tokens equally is regarded as a low-level global feature \( v_l \in \mathbb{R}^d \). The above features constitute a multi-level global image feature set \( V^S = \{ v_l, v_m, v_h \} \in \mathbb{R}^{3 \times d} \). Similarly, a similar process for text \( T \) is performed to generate a multi-level global text feature set \( T^S = \{ t_l, t_m, t_h \} \in \mathbb{R}^{3 \times d} \).

D. Cross-Grained Feature Refinement

In the MGF module, our focus is on mining fine-grained discriminative information using selected informative tokens within each modality. The reliability of these selected informative tokens is crucial. As TIREID aims for cross-modal alignment, it is desirable for the selected informative tokens to be both highly discriminative and modality-shared. The selection process relies on the attention response score between the class token and other tokens, which is actually determined by the discrimination ability of the class token. However, despite the discriminative capacity of our network initialized by CLIP, the sensitivity of CLIP to fine-grained information limits the discriminativeness of the class tokens generated by the initial network. Consequently, some non-modality-shared tokens may be selected, posing challenges for model optimization and potentially leading to trivial solutions. To address this challenge, we introduce a cross-grained (i.e. image-word and text-patch) feature refinement (CFR) module to suppress and filter out non-modality-shared information in the selected tokens. Since token selection is based on the correlation between each local token and class token, we filter out unimportant information by the similarity between local and global features across modalities. The CFR is illustrated in Figure 4(bottom).

Given the image-level representation \( v_g \in \mathbb{R}^d \), the selected patch-level representation \( V^S \in \mathbb{R}^{2K_t \times d} \), the sentence-level representation \( t_S \in \mathbb{R}^d \) and the selected word-level representation \( T^S \in \mathbb{R}^{2K_t \times d} \). The image-word and sentence-patch similarities are evaluated by inner products between corresponding feature representations, respectively.

\[
\begin{align*}
S_{I-W} &= ( T^S v_g )^T \\
S_{P-S} &= V^T t_S
\end{align*}
\]  

where \( S_{I-W} \in \mathbb{R}^{1 \times 2K_t} \) is the similarity between the image and the selected words in the sentence, and \( S_{P-S} \in \mathbb{R}^{2K_t \times 1} \) is the similarity between the sentence and the selected patches of an image. Then we fuse the above similarities to get the instance-level similarity. To emphasize important information and filter out non-modality-shared information in the selected image and text tokens, we adaptively generate different weights for each score in \( S_{I-W} \) and \( S_{P-S} \) by Softmax during aggregation, where the scores for the image patches (words) related to the sentence (image) will be given high weights. Finally, the instance-level similarity is generated by computing a weighted sum of each score in \( S_{I-W} \) and \( S_{P-S} \).

\[
\begin{align*}
S_{I-W} &= \sum_{i=1}^{2K_t} \frac{2K_t}{\sum_{j=1}^{2K_t} \exp ( S_{I-W}(1,i) )} S_{I-W}(1,i) \\
S_{P-S} &= \sum_{i=1}^{2K_t} \frac{2K_t}{\sum_{j=1}^{2K_t} \exp ( S_{P-S}(1,i) )} S_{P-S}(1,i)
\end{align*}
\]  

Fig. 3. (a) Illustration of the token selection process and (b) structure of multi-level global feature learning (MGF) module, where \([CLS] ([CLS]^V, [CLS]^T])\) is the learnable class token, \(K(K_t/K_e)\) denotes the number of informative tokens, \(F(S^V/T^S)\) represents the multi-level global feature set.
the feature space. However, individual words and patches may have vague meanings, and each word (patch) may be related to multiple patches (words), as shown in Figure 1. Accordingly, we pick out the most related \( K_p \) positive words (image patches) for each image patch (word) to construct the patch-word (word-patch) pairs. Furthermore, to optimize computational efficiency, we focus on the most informative image patches and words for the correspondence discovery process.

Formally, for image-text pair \((I, T)\), the most informative patches and words are \( V_h^t \in \mathbb{R}^{K_v \times d} \) and \( T_h^t \in \mathbb{R}^{K_t \times d} \), respectively. We first compute the cosine similarity between \( V_h^t \) and \( T_h^t \). For example, the first row \((T_h^t V_h^t)_{(1,:)} \in \mathbb{R}^{K_t}\) of the similarity matrix denotes dependency of the \(1^{th}\) word to each informative patch. Then, the Topk operation is used to select the most related \( K_p \) patches for the \(1^{th}\) word \( t^1\). Finally, these \( K_p \) patches are average pooled to generate a new patch feature \( v^1_p \), and \((t^1_v, v^1_p)\) forms a matched word-patch pair.

\[
\mathbf{w}^1 = \text{AvgPool}(\text{Topk}(\text{Softmax}(T_h^t V_h^t)_{(1,:)})) \quad (9)
\]

And so on, we can generate \( K_h \) matched patch-word pairs \((v^1_p, t^1_v), \ldots, (v^K_p, t^K_v)\) and \( K_t \) matched word-patch pairs \((t^1_i, v^K_i), \ldots, (t^K_{i}, v^K_{i})\). To make these matched pairs close to each other in the feature space, we get the instance-level similarity by computing the sum of the similarities between all \( K_h + K_t \) matched pairs.

\[
s_f(I, T) = \sum_{i=1}^{K_h} s(v^i_p, t^i_v) + \sum_{j=1}^{K_t} s(v^j_p, t^j_v) \quad (10)
\]

where \( s(\cdot) \) denotes the cosine similarity metric, and \( s_f \) denotes the similarity between image \( I \) and text \( T \) obtained by the FCD module. Similarly, we also maximize the similarity \( s_f \) to constrain the reliability of selected tokens at a finer level.

**E. Fine-Grained Correspondence Discovery**

The CFR module enforces coarse-level constraints on the reliability of selected tokens through global-local matching between modalities. In this section, we propose a fine-grained correspondence discovery (FCD) module that enforces finer-level constraints by mining local-local fine-grained correspondences of selected tokens between images and texts, as shown in Figure 4(top). The simplest way to mine the fine-grained correspondences is to find out the most related positive word (image patch) as the correspondence for each image patch (word) to construct the patch-word (word-patch) pairs, and make them close to each other in the feature space.

\[
s_c(I, T) = (s_{I-W} + s_{P-S})/2 \quad (8)
\]

where \( s_c \in \mathbb{R}^1 \) denotes the similarity between image \( I \) and text \( T \) after cross-grained feature refinement. Since \( I \) and \( T \) are matched, we maximize the similarity \( s_c \) to ensure that the selected tokens are as modality-shared as possible.

**F. Training and Inference**

The commonly used cross-modal projection matching (CMPM) loss and cross-modal projection classification (CMPC) loss proposed by [24] is adopted as our training objective function to learn image-text alignment.

1) **Cross-Modality Alignment**: For image \( I \) and text \( T \), we generate multi-level global feature sets \( \{I_l, v_m, v_k\} \) and \( \{t_l, t_m, t_h\} \) by MGF, respectively. To align \( I \) and \( T \), we compute CMPM and CMPC losses for each level of image and text features to supervise the multi-level global feature learning.

\[
\mathcal{L}_{cm} = \sum_{i \in \{l,m,h\}} (\mathcal{L}^i_{\text{cmpm}} + \mathcal{L}^i_{\text{cmpc}}) \quad (11)
\]

Besides, since \( s_c \) and \( s_f \) both represent the similarity between \( I \) and \( T \), we sum them to represent the final similarity \( s_{cf} \). For a batch of \( B \) image-text pairs, we expect to maximize the similarity if the image and text are matched and minimize it otherwise. Thus, a cross-modality bi-directional dual-constrained triplet ranking loss is used to optimize it.

\[
\begin{align*}
\mathcal{L}_c &= \max(\alpha - s_{cf}(I, T) + s_{cf}(I, T_n), 0) \\
&\quad + \max(\alpha - s_{cf}(I, T) + s_{cf}(I_n, T), 0)
\end{align*}
\quad (12)
\]
where \((I, T)\) denotes the matched image-text pairs, and \((I, T_n)\), \((I_n, T)\) denote the mismatched pairs. \(T_n\) and \(I_n\) denote the hard negative samples. \(\alpha\) indicates the margin.

2) Diversity Regularization: To fully mine fine-grained details in images and texts, it is desirable for different-level features to emphasize distinct information. To this end, we impose a diversity constraint \(\mathcal{L}_{div}\) on different-level features, avoiding information redundancy, which is represented as follows:

\[
\mathcal{L}_d = \sum_{i \in \{l, m, h\}} \sum_{j \in \{l, m, h\}, i \neq j} \left( \frac{v_i \cdot v_j}{\|v_i\|_2 \|v_j\|_2} + \frac{t_i \cdot t_j}{\|t_i\|_2 \|t_j\|_2} \right)
\]

3) Objective Function: Integrating the above constraints, the final objection function \(\mathcal{L}\) is as follows.

\[
\mathcal{L} = \mathcal{L}_{cm} + \lambda_c \mathcal{L}_c + \lambda_d \mathcal{L}_d
\]

where \(\lambda_c\) and \(\lambda_d\) balance the focus on different loss terms during training.

4) Inference: Note that the CFR and FCD modules are only utilized for training and are removed during inference to optimize computational efficiency. During inference, for the text query and image candidate, we generate the multi-level global text and image features. The similarity between the text-image pair is then computed as the sum of the cosine distances of different-level features, i.e., \(s = s_l + s_m + s_h\).

IV. EXPERIMENTS

A. Experiment Settings

1) Datasets and Metrics: **CUHK-PEDES** [6] is previously the only accessible large-scale benchmark for TIReID. It includes 40,206 images and 80,412 text descriptions of 13,003 persons. Each image is manually annotated with 2 descriptions, with an average length of not less than 23 words. Follow [6], 34,054 images and 68,108 descriptions of 11,003 persons, 3,078 images and 6,156 descriptions of 1000 persons, 3,074 images and 6,148 descriptions of 1000 persons are utilized for training, validation, and testing, respectively. Recently, several large-scale datasets [10], [11] have been released, which greatly promoted the development of TIReID. **ICFG-PEDES** [11] contains 54522 text descriptions for 54,522 images of 4,102 persons. Each description has an average length of 37 words, and the vocabulary contains 5554 unique words. Compare with CUHK-PEDES, the text description of ICFG-PEDES is more identity-centric and fine-grained. The dataset is split into train, and test with 34674 image-text pairs of 3102 persons, and 19848 image-text pairs of the remaining 1000 persons, respectively. **RST-P Reid** [10] is constructed to handle real scenarios. It includes 41010 textual descriptions and 20505 images of 4101 persons. Specifically, each person contains 5 images caught by 15 cameras. Each image corresponds to 2 text descriptions, and the length of each description is no shorter than 23 words. The dataset is split into 3701 train, 200 validation, and 200 test persons. To verify the effectiveness of our method, we conduct extensive experiments on the above three benchmarks. We adopt the Recall at Rank K (Rank-K, higher is better) as the retrieval metric to evaluate the retrieval performance, where the Rank-1, Rank-5, and Rank-10 accuracy are reported.

2) Implementation Details: We conduct our experiments using the PyTorch library on a single RTX3090 24GB GPU. Input images are resized to 224 × 224, and random horizontal flipping is applied for data augmentation. The maximum text length for all datasets was set to 100. The length of visual and textual token sequences after tokenization is \(N_v = 196\) and \(N_t = 100\). The dimension \(d\) of image and text embeddings is 768. The MGF module consists of \(M = 1\) GLD block, with each block having 12 heads. Besides, \(2R_v = 20\%\) informative image tokens and \(2R_t = 40\%\) informative text tokens in MGF are selected to learn multi-level global discriminative features. In the FCD module, we pick out \(K_p = 3\) most related words (patches) for each patch (word) to construct the patch-word (word-patch) pair. The loss balance factors are set to \(\lambda_c = 10\) and \(\lambda_d = 0.2\). The margin \(\alpha\) in triplet loss is set to 0.2. During training, our model is optimized using Adam optimizer with a liner warmup strategy. We employ different learning rates for different modules: the initial learning rate for the image and text backbone is set to 1e-5, while other modules of the network are initialized to 1e-4. The learning rate is decreased by a factor of 0.1 at the 20th, 25th, and 35th epoch. The network is trained with a batch size of 32 for a total of 50 epochs.

B. Comparisons With State-of-the-Art Models

In this section, we evaluate our proposed CFine under two different settings: using ViT pre-trained on ImageNet (Ours-IMG) or using ViT from CLIP (Ours-CLIP) as the image encoder. We compare our method with state-of-the-art approaches on three standard TIReID benchmarks, and the results are summarized in Tables I, II and III. Our method consistently achieves state-of-the-art results on all three benchmarks with significant improvements.

1) **CUHK-PEDES**: We first evaluate our CFine on the widely-used benchmark, **CUHK-PEDES**, and the performance comparison is shown in Table I. Remarkably, CFine consistently achieves state-of-the-art results under both settings. When using ViT pre-trained on ImageNet as the image encoder, CFine reaches remarkable Rank-1, Rank-5, and Rank-10 accuracy of 65.07%, 83.01% and 89.00%, respectively, outperforming the Transformer-based methods IVT [30] and SAF [31]. Furthermore, CFine surpasses the recent state-of-the-art method AXM-Net [64], showcasing the effectiveness of our proposed fine-grained information excavation modules (MGF, CFR, and FCD) in reducing the modality gap. It is worth mentioning that AXM-Net requires computationally expensive cross-modality interaction operations. With ViT from CLIP as the image encoder, CFine demonstrates even more substantial performance improvements compared to all the methods. Particularly, when compared with the strongest competitor AXM-Net [64], CFine achieves 69.57% (+5.13%), 85.93% (+5.41%) and 91.15% (+4.38%) of Rank-1, Rank-5 and Rank-10 accuracy, respectively. This highlights the benefits of introducing ample cross-modal correspondence prior to our approach. However, TextReID [27] also introduces CLIP to facilitate learning, and its performance is even inferior to some non-CLIP methods, which is attributed to its failure to effectively bridge the task gap that CLIP is not sensitive
to fine-grained information. While our proposed fine-grained information excavation modules bridge the task gap well.

2) Other Benchmarks: To assess the generalization of our method, we conduct comparisons with previous works on two additional benchmarks, namely ICFG-PEDES and RSTPReid, as presented in Tables II and III. The results demonstrate the competitive performance of our CFine on ICFG-PEDES and RSTPReid when using ViT pre-trained on ImageNet as the image encoder, outperforming all methods except IVT [30]. When using ViT from CLIP as the image encoder, CFine outperforms all existing methods by a large margin across all metrics. For instance, CFine achieves a Rank-1 accuracy of 60.83% on ICFG-PEDES and 50.55% on RSTPReid, surpassing IVT [30] on Rank-1 by +4.79%, +3.85% respectively. These results demonstrate the generalization and robustness of our proposed CFine.

The above results and analysis show that our CFine achieves consistent improvements across all benchmarks. These improvements can be attributed to several key factors.

1) our proposed fine-grained information excavation modules can fully mine modality-shared fine-grained discriminative details within each modality to enhance global features, effectively narrowing the modality gap and distinguishing different pedestrians. (2) Introducing ample cross-modal correspondences contained in CLIP can bring significant performance gains. (3) The incorporation of fine-grained information excavation enables the successful transfer of the cross-modal representation capacity learned from the upstream VLP task to the TIReID task.

C. Ablation Studies

To comprehensively evaluate the impact of different modules in CFine, we conduct extensive ablation studies to compare various variants of CFine on CUHK-PEDES. For these experiments, we employed ViT from CLIP as the image encoder. To be specific, we systematically analyze the contribution of each component in CFine by combining different components. Furthermore, we discuss the important parameters and variants within each module to gain insights into their individual effects on performance.

1) Contributions of Algorithmic Components: We examine the contributions of each module in Table IV. No.0 shows the results of Baseline. Baseline means only using ViT pre-trained on ImageNet and the language pre-trained model BERT as the image and text encoders to extract features without adding any modules. Comparing the results of No.0 and No.1 in Table IV, we observe that when replacing the image encoder of Baseline with ViT from CLIP, the performance increases by 7.67%, 5.28%, 4.08% respectively. These results demonstrate the generalization and robustness of our proposed CFine.
TABLE IV

ABSTRACTION STUDY ON DIFFERENT COMPONENTS OF OUR PROPOSED CFine ON CUHK-PEDES

| No. | Methods | CLIP | MGF | CFR | FCD | Rank-1 | Rank-5 | Rank-10 | Params | FLOPs |
|-----|---------|------|-----|-----|-----|--------|--------|--------|--------|-------|
| 0   | Baseline |      |     |     |     | 57.89  | 78.46  | 85.77  | 195.28M | 26.269 |
| 1   | +CLIP   | ✓    |     |     |     | 65.56  | 83.74  | 89.85  | 195.68M | 26.270 |
| 2   | +CLIP+MGF (m) | ✓ | ✓    |     |     | 67.66  | 85.10  | 89.80  | 204.74M | 26.970 |
| 3   | +CLIP+MGF (h) | ✓ | ✓    | ✓    |     | 68.22  | 85.12  | 90.24  | 204.74M | 26.970 |
| 4   | +CLIP+CFR | ✓ | ✓    |     | ✓    | 66.41  | 84.16  | 89.91  | 195.28M | 26.272 |
| 5   | +CLIP+FCR | ✓ | ✓    |     | ✓    | 66.13  | 84.15  | 89.78  | 195.28M | 26.289 |
| 6   | +CLIP+MGF | ✓ | ✓    | ✓    | ✓    | 68.62  | 85.36  | 90.84  | 204.74M | 27.670 |
| 7   | +CLIP+MGF+CFR | ✓ | ✓    | ✓    | ✓    | 69.14  | 85.40  | 90.55  | 204.74M | 27.672 |
| 8   | +CLIP+MGF+FCD | ✓ | ✓    | ✓    | ✓    | 69.33  | 85.22  | 90.61  | 204.74M | 27.689 |
| 9   | CFine   | ✓ | ✓    | ✓    | ✓    | 69.57  | 85.93  | 91.15  | 204.74M | 28.691 |

into the model. The results of No.1 vs No.2 and No.1 vs No.3 reveal the efficacy of the MGF module. When adding MGF (m) or MGF (h) to No.1, the Rank-1 accuracy improves by 2.1% or 2.66%, respectively. Besides, as shown in the result of No.6, the combination of MGF (m) and MGF (h) can further improve performance, resulting in a boost in Rank-1 accuracy from 65.56% to 68.62%. These results in No.2, No.3, and No.6 justify that MGF effectively mines the discriminative local clues and learns discriminative global features at multiple levels.

The CFR module is used to filter redundant information and ensure the reliability of informative tokens at a coarser level, while the FCD module is employed to fully discover inter-modal fine-grained correspondences and enforce constraints on the reliability of informative tokens at a finer level. These results of No.1 vs No.4, No.1 vs No.5, No.6 vs No.7, and No.6 vs No.8 demonstrate the effectiveness of CFR and FCD. When all modules of CFine are combined, the best retrieval performance is achieved. Furthermore, comparing the results of No.1 and No.9 reveals that the direct usage of CLIP can be sub-optimal for TIReID due to the substantial gap between instance-level pre-training and fine-grained TIReID tasks. However, when CLIP is integrated with our proposed fine-grained information excavation modules (MGF, CFR, and FCD), the performance is substantially improved from 65.56% to 69.57%. This result demonstrates that the ample cross-modal knowledge of CLIP is fully exploited and transferred to the TIReID task through our proposed modules.

2) Ablation of Multi-Level Global Feature Learning: In the MGF module, the token selection first selects the most informative image and text tokens, and then feeds them into GLD consisting of \( M \) blocks to learn a set of multi-level global features. We conducted an analysis to examine the impact of several crucial factors in MGF, namely the image token selection ratio \( R_v \), text token selection ratio \( R_t \), and the number of GLD blocks \( M \). The results of this analysis are presented in Figure 5(a)(b)(c). As shown in the figure, both \( R_v \) and \( R_t \) have a noticeable influence on performance. If the ratio is too small, it fails to contain comprehensive discriminative local information, resulting in information loss. If the ratio is too large, noise will be introduced, which will have a negative impact on the model. Therefore, we set \( R_v = 0.1 \) and \( R_t = 0.2 \). Regarding the number of GLD blocks \( M \), the results indicate that \( M = 1 \) yields the best performance.

We conduct additional experiments to compare MGF with other variants, as summarized in Table V. In order to verify the effectiveness of the multi-level learning way, we attempt feeding \( 2K_v/2K_t \) tokens into GLD at once (MGF-2K) to learn discriminative global features. The result shows severe performance degradation (-2.14%). We attribute this to the fact that the one-time learning way can overwrite the role of crucial informative tokens, hindering the extraction of discriminative

| No. | Methods | CLIP | MGF | CFR | FCD | Rank-1 | Rank-5 | Rank-10 | Params | FLOPs |
|-----|---------|------|-----|-----|-----|--------|--------|--------|--------|-------|
| 0   | Baseline |      |     |     |     | 57.89  | 78.46  | 85.77  | 195.28M | 26.269 |
| 1   | +CLIP   | ✓    |     |     |     | 65.56  | 83.74  | 89.85  | 195.68M | 26.270 |
| 2   | +CLIP+MGF (m) | ✓ | ✓    |     |     | 67.66  | 85.10  | 89.80  | 204.74M | 26.970 |
| 3   | +CLIP+MGF (h) | ✓ | ✓    | ✓    |     | 68.22  | 85.12  | 90.24  | 204.74M | 26.970 |
| 4   | +CLIP+CFR | ✓ | ✓    |     | ✓    | 66.41  | 84.16  | 89.91  | 195.28M | 26.272 |
| 5   | +CLIP+FCR | ✓ | ✓    |     | ✓    | 66.13  | 84.15  | 89.78  | 195.28M | 26.289 |
| 6   | +CLIP+MGF | ✓ | ✓    | ✓    | ✓    | 68.62  | 85.36  | 90.84  | 204.74M | 27.670 |
| 7   | +CLIP+MGF+CFR | ✓ | ✓    | ✓    | ✓    | 69.14  | 85.40  | 90.55  | 204.74M | 27.672 |
| 8   | +CLIP+MGF+FCD | ✓ | ✓    | ✓    | ✓    | 69.33  | 85.22  | 90.61  | 204.74M | 27.689 |
| 9   | CFine   | ✓ | ✓    | ✓    | ✓    | 69.57  | 85.93  | 91.15  | 204.74M | 28.691 |

Fig. 5. Effect of different parameters, including (a) the selected ratio \( R_v \) of image tokens (top left); (b) the selected ratio \( R_t \) of text tokens (top right); (c) the amount \( M \) of GLD blocks (bottom left); (d) the number \( K_p \) of selected positive patches (words) for each word (patch) (bottom right).
fine-grained clues. Our multi-level learning way allows for the comprehensive mining of such clues. In GLD, we conduct the cross-attention between selected tokens and all tokens to learn global features. Another possible choice is to compute the self-attention between selected tokens. Our scheme achieves superior performance, which mainly attributes to the fact that our scheme leverages information not only from informative tokens but also from other contexts. In addition, before being fed into GLD, the high-level and middle-level token sequences are padded with $[CLS_h]$ and $[CLS_m]$ at the beginning to learn multi-level global features, respectively. The initialization of these $[CLS]$ tokens is critical. We compare three different initialization methods: (1) initialization using the mean value of the selected token sequences ($CLS$-mean); (2) random initialization ($CLS$-random); (3) initialization using the global features output by the encoder (Ours). The results show that the third initialization achieves the best performance. We believe that initializing $[CLS]$ with the output of the encoder strengthens the connection between the GLD module and the encoder, fostering their collaboration and further emphasizing the important fine-grained clues based on the output of the encoder.

3) Ablation of Cross-Grained Feature Refinement: In the CFR module, each score in image-word $S_{I-W}$ and sentence-patch $S_{P-S}$ similarities needs to be aggregated to form the instance-level similarity for the next text-image matching, in which score aggregation strategy is crucial. We compared different aggregation strategies, as shown in Table VI. For the $2K_p$ ($2K_p$) scores in image-word (sentence-patch) similarity, we get the instance-level score by the following strategies: (1) Summing all scores (Aggr-Sum); (2) Taking the mean of all scores (Aggr-Mean); (3) Taking the maximum score (Aggr-Max); (4) Computing a weighted sum of all scores, with weights generated by Softmax (Ours). The results clearly indicate that our aggregation strategy achieves the best results. This can be attributed to the weighting way that filters out the redundant information and highlights the important information. Notably, our CFR module characterizes sample-pair similarity by directly computing the cross-grained feature similarity between modalities, without introducing additional parameters. To further demonstrate the advantage of the CFR module, we follow [66] by introducing a learnable multimodal encoder (MME) to predict the similarity scores of sample pairs. Specifically, we first feed the cross-grained feature sequence between modalities into self-attention (MME-SA) or cross-attention (MME-CA) modules to interact and generate a fused multimodal feature. We then pass the feature through a feed-forward network to predict the similarity score of the sample pair. Tables VI and VII show that the performance of MME is much lower than our CFR, which introduces more uncertainties and increases the difficulty of model optimization. Additionally, MME also incurs a significant increase in computational cost. In contrast, our CFR module achieves superior performance without introducing extra parameters. This further confirms the superiority of our CFR module.

4) Ablation of Fine-Grained Correspondence Discovery: The FCD module picks out the most related words (patches) for each patch (word) to discover fine-grained correspondence. We conduct experiments to determine the optimal value of $K_p$ for performance, as shown in Figure 5(d). The results indicate that If $K_p$ is too small, the meaning will be ambiguous, and the network excessively focuses on local correspondence, leading to over-fitting. Conversely, if $K_p$ is too large, irrelevant information is introduced, preventing the establishment of accurate correspondence. Remarkably, When $K_p = 3$, our proposed method achieves the best results.

5) Computational Complexity: In this section, we analyze the computational complexity of our proposed method, reporting the number of model parameters (Params), the number of floating-point operations for an input image (FLOPs) during training, and retrieval time (Time) at the inference stage. Table IV presents the Params and FLOPs for several variants of our model. Compared with Baseline, only the MGF module introduces a slight increase in computational cost, while the CFR and FCD modules do not add any extra parameters and exhibit almost no increase in FLOPs. We further compare our method with several typical local-matching TIREID methods, such as SSAN [11], TIPCB [12], and SAF [31]. As depicted in Table VII, SSAN has a low number of parameters due to its ResNet50+LSTM structure, but the inclusion of Non-Local attention leads to relatively high FLOPs. TIPCB, based on the structure of ResNet50+TextCNN, relies on numerous residual blocks to learn local features, resulting in a significant increase in computational cost. Both SAF and our CFine are Transformer-based (ViT+BERT) networks with a higher number of parameters. While SAF introduces an additional multi-head attention module to generate local features, it incurs the highest computation and retrieval time. Due to the presence of self-attention in Transformer, Baseline requires more retrieval time compared to other CNN-based methods. However, when compared with Baseline, CFine only increases the retrieval time by 3.5s, thus confirming the high efficiency.

### Table VI

| Method   | Rank-1 | Rank-5 | Rank-10 |
|----------|--------|--------|---------|
| Aggr-Sum | 69.07  | 85.45  | 90.72   |
| Aggr-Mean| 68.66  | 85.32  | 90.46   |
| Aggr-Max | 68.44  | 85.15  | 90.11   |
| MME-SA   | 68.88  | 85.53  | 90.51   |
| MME-CA   | 68.79  | 85.69  | 90.56   |
| Ours     | 69.57  | 85.93  | 91.15   |

### Table VII

| Method | Params | FLOPs  | Time   | Rank-1 |
|--------|--------|--------|--------|--------|
| Baseline | 195.28M | 26.269 | 35.7s  | 57.89  |
| MME-SA  | 211.83M | 45.987 | 57.2s  | 68.88  |
| MME-CA  | 211.83M | 31.015 | 57.2s  | 68.79  |
| SSAN [11] | 97.86M | 18.139 | 21.4s  | 61.37  |
| TIPCB [12] | 184.75M | 43.861 | 25.1s  | 64.26  |
| SAF [31] | 224.78M | 32.276 | 43.5s  | 64.13  |
| Ours    | 204.74M | 27.691 | 37.2s  | 69.57  |
of our designed modules. Overall, our method demonstrates competitive computational cost and retrieval efficiency when compared to existing approaches.

6) Qualitative Results: In this section, we qualitatively verify the effectiveness of each module in our proposed method. Figure 6 provides some visualization examples of the image-word and sentence-patch scores obtained from the CFR module. In each image-text pair, the first row displays the similarity score map between the image feature and the selected $2K_t$ (40) informative word features. While the second row shows the similarity score map between the sentence feature and the selected $2K_v$ (38) informative patch features. For ease of presentation, each score map is condensed into two lines: the first line represents the scores corresponding to the word (patch) feature from Top-1 to Top-$K_t$ (Top-$K_v$), and the second line represents the scores corresponding to the word (patch) feature from Top-$K_t$+1 (Top-$K_v$+1) to Top-2$K_t$ (Top-2$K_v$). The darker the score map, the greater the similarity score. As observed in Figure 6, the words/patches selected by our network are mostly discriminative and modality-shared. While the network may occasionally select some unimportant words/patches, the corresponding scores assigned to these words/patches are relatively small, and the network assigns lower weights to these scores to suppress their impact.

Figure 7 presents a comparison of attention response maps among various variants of CFine. These attention response maps reflect discriminative fine-grained information emphasized by global features, with brighter regions indicating stronger responses to corresponding local details. Figure 7 shows that Baseline (the second column) that only performs simple global alignment has a high response to the entire image, lacking the ability to focus on discriminative information. In the third column (CLIP), the network benefits from the powerful image-text matching ability of CLIP and pays more attention to the effective human body area, yet it still lacks sufficient attention to fine-grained information. The sixth, seventh, and eighth columns (CLIP+MGF) depict attention response maps corresponding to low-level, mid-level, and high-level global features, respectively. Introducing our MGF module enables the network to fully mine fine-grained discriminative information, particularly evident in the mid-level and high-level attention response maps. Furthermore, different-level features emphasize different discriminative information, and the combination of multi-level features allows for a more comprehensive mining of fine-grained information. However, upon examining the response maps of Baseline, CLIP, and CLIP+MGF, it becomes evident that the network pays attention to non-modality-shared information, especially the image background. When our CFR and FCD modules are introduced (CFine), the network shifts its focus towards mining modality-shared fine-grained discriminative information. This confirms the advantages of the CFR and FCD modules in ensuring the reliability of informative tokens. These visualization results qualitatively validate the motivation and effectiveness of our modules.

Figure 8 shows the top-10 retrieval results obtained using Baseline+CLIP and CFine for the given text query. The difference between the two lies in the incorporation of fine-grained information excavation. From the figure, it is
In this work, we present CFine, a novel transformer architecture with fine-grained information excavation. Our goal is to leverage the capabilities of CLIP to achieve cross-modal fine-grained alignment for TIREID. To take full advantage of the rich multi-modal knowledge from CLIP, we performed fine-grained information excavation. Specifically, the MGF module enhances intra-modal fine-grained discriminative clues by modeling interaction between global image (text) and local discriminative patches (words). By modeling cross-grained information excavation. Specifically, the MGF module filters out non-modality-shared local information at a coarser level. The FCD module establishes the fine-grained cross-modal correspondence by capturing the relationship between image patches and words. These modules work synergistically to effectively transfer the knowledge of the CLIP model to TIREID. Significant performance gains on three popular TIREID benchmarks prove the superiority and effectiveness of the proposed CFine.

V. CONCLUSION

In this work, we present CFine, a novel transformer architecture with fine-grained information excavation. Our goal is to leverage the capabilities of CLIP to achieve cross-modal fine-grained alignment for TIREID. To take full advantage of the rich multi-modal knowledge from CLIP, we performed fine-grained information excavation. Specifically, the MGF module enhances intra-modal fine-grained discriminative clues by modeling interaction between global image (text) and local discriminative patches (words). By modeling cross-grained information excavation. Specifically, the FCD module establishes the fine-grained cross-modal correspondence by capturing the relationship between image patches and words. These modules work synergistically to effectively transfer the knowledge of the CLIP model to TIREID. Significant performance gains on three popular TIREID benchmarks prove the superiority and effectiveness of the proposed CFine.

REFERENCES

[1] L. Zheng, Y. Huang, H. Lu, and Y. Yang, “Pose-invariant embedding for deep person re-identification,” IEEE Trans. Image Process., vol. 28, no. 9, pp. 4500–4509, Sep. 2019.
[2] H. Li, S. Yan, Z. Yu, and D. Tao, “Attribute-identity embedding and self-supervised learning for scalable person re-identification,” IEEE Trans. Circuits Syst. Video Technol., vol. 30, no. 10, pp. 3472–3485, Oct. 2020.
[3] S. Yan, Y. Zhang, M. Xie, D. Zhang, and Z. Yu, “Cross-domain person re-identification with pose-invariant feature decomposition and hyper-graph structure alignment,” Neurocomputing, vol. 467, pp. 229–241, Jan. 2022.
[4] Y. Wang, J. Peng, H. Wang, and M. Wang, “Progressive learning with multi-scale attention network for cross-domain vehicle re-identification,” Sci. China Inf. Sci., vol. 65, no. 6, Jun. 2022, Art. no. 160103.
[5] L. Zhang, G. Du, F. Liu, H. Tu, and X. Shu, “Global-local multiple granularity learning for cross-modality visible-infrared person re-identification,” IEEE Trans. Neural Netw. Learn. Syst., early access, Jun. 17, 2021, doi: 10.1109/TNNLS.2021.3085978.
[6] S. Li, T. Xiao, H. Li, B. Zhou, D. Yue, and X. Wang, “Person search with natural language description,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 5187–5196.
[7] Z. Wang, Z. Fang, J. Wang, and Y. Yang, “VITAA: Visual-textual attributes alignment in person search by natural language,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2022, pp. 2606–2614.
[8] S. Gaggarwal, R. V. Babu, and A. Chakraborthy, “Text-based person search via attribute-aided matching,” in Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV), Mar. 2020, pp. 2606–2614.
[9] C. Gao et al., “Contextual non-local alignment over full-scale representation for text-based person search,” 2021, arXiv:2101.03056.
[10] A. Zhu et al., “DSSL: Deep surroundings-person separation learning for text-based person retrieval,” in Proc. ACM Int. Conf. Multimedia (MM), 2021, pp. 209–217.
[11] Z. Ding, C. Ding, Z. Shao, and D. Tao, “Semantically self-aligned network for text-to-image part-aware person re-identification,” 2021, arXiv:2107.12666.
[12] Y. Chen, G. Zhang, Y. Lu, Z. Wang, and Y. Zheng, “TIPCB: A simple but effective part-based convolutional baseline for text-based person search,” Neurocomputing, vol. 494, pp. 171–181, Jul. 2022.
[13] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 770–778.
[14] A. Dosovitskiy et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” in Proc. Int. Conf. Learn. Represent. (ICLR), 2021, pp. 1–9.
[15] J. Devlin, M. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in Proc. North Amer. Chapter Assoc. Comput. Linguistics (NAACL), 2019, pp. 4171–4186.
[16] A. Radford et al., “Learning transferable visual models from natural language supervision,” in Proc. Int. Conf. Mach. Learn. (ICML), 2021, pp. 8748–8763.
[17] K. Zheng, W. Liu, J. Liu, Z.-J. Zha, and T. Mei, “Hierarchical Gumbel attention network for text-based person search,” in Proc. 28th ACM Int. Conf. Multimedia, Oct. 2020, pp. 1–6.
[18] K. Niu, Y. Huang, W. Ouyang, and L. Wang, “Improving description-based person re-identification by multi-granularity image-text alignments,” IEEE Trans. Image Process., vol. 29, pp. 5542–5556, 2020.
[19] W. Zhao, L. Zhang, Y. Yang, Q. Tian, and S. Wang, “Beyond part models: Person retrieval with refined part pooling (and a strong convolutional Baseline),” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2018, pp. 480–496.
[20] Y. Jing, C. Si, J. Wang, W. Wang, L. Wang, and T. Tan, “Pose-guided multi-granularity attention network for text-based person search,” in Proc. AAAI Conf. Artif. Intell. (AAAI), vol. 34, no. 7, 2020, pp. 11189–11196.
[21] K. Lee, X. Chen, G. Hua, H. Hu, and X. He, “Stacked cross attention for image-text matching,” in Eur. Conf. Comput. Vis., (ECCV), 2018, pp. 1–16.
[22] C. Wang, Z. Luo, Y. Lin, and S. Li, “Text-based person search via multi-granularity embedding learning,” in Proc. 13th Int. Joint Conf. Artif. Intell., Aug. 2021, pp. 1–9.

[23] Z. Zheng, L. Zheng, M. Garrette, Y. Yang, M. Xu, and Y. Shen, “Dual-path convolutional image-text embeddings with instance loss,” ACM Trans. Multimedia Comput., Commun., Appl., vol. 16, no. 2, p. 51, 2020.

[24] Y. Zhang and H. Lu, “Deep cross-modal projection learning for image-text matching,” in Proc. Eur. Conf. Comput. Vis., (ECCV), 2018, pp. 1–9.

[25] N. Sarafianos, X. Xu, and I. Kakadiaris, “Adversarial representation learning for text-to-image matching,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 5813–5823.

[26] Y. Chen, R. Huang, H. Chang, C. Tan, T. Xue, and B. Ma, “Cross-modal knowledge adaptation for language-based person search,” IEEE Trans. Image Process., vol. 30, pp. 4057–4069, 2021.

[27] X. Han, S. He, L. Zhang, and T. Xiang, “Text-based person search with limited data,” in Proc. Brit. Mach. Vis. Conf. (BMVC), 2021, pp. 1–13.

[28] Y. Wu, Z. Yan, X. Han, G. Li, C. Zou, and S. Cui, “LapsCore: Language-guided person search via color reasoning,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2021, pp. 1604–1613.

[29] S. Yan, H. Tang, L. Zhang, and J. Tang, “Image-specific information suppression and implicit local alignment for text-based person search,” 2022, arXiv:2208.14365.

[30] X. Shu et al., “See finer, see more: Implicit modality alignment for text-based person retrieval,” in Proc. Eur. Conf. Comput. Vis. Workshop Real World Surveill. (ECCVW), 2022, pp. 624–643.

[31] S. Li, M. Cao, and M. Zhang, “Learning semantic-aligned feature representation for text-based person search,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), May 2022, pp. 2724–2728.

[32] Z. Shao, X. Zhang, M. Fang, Z. Lin, J. Wang, and C. Ding, “Learning granularity-unified representations for text-to-image person re-identification,” in Proc. 30th ACM Int. Conf. Multimedia, Oct. 2022, pp. 1–9.

[33] G. Jia et al., “Scaling up visual and vision-language representation learning with noisy text supervision,” in Proc. Int. Conf. Mach. Learn. (ICML), 2021, pp. 1–9.

[34] L. Yao et al., “FILIP: Fine-grained interactive language-image pre-training,” in Proc. Int. Conf. Learn. Represent. (ICLR), 2022.

[35] M. Cao, T. Yang, J. Weng, C. Zhang, J. Wang, and Y. Zou, “LocVTP: Video-text pre-training for temporal localization,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2022, pp. 38–53.

[36] L. Harold Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, “VisualBERT: A simple and performant baseline for vision and language,” 2019, arXiv:1908.01557.

[37] Y. Chen et al., “UNITER: Universal image-text representation learning,” in Proc. Eur. Conf. Comput. Vis. ECCV, 2020, pp. 104–120.

[38] H. Luo et al., “CLIP4Clip: An empirical study of CLIP for end to end video clip retrieval and captioning,” Neurocomputing, vol. 508, pp. 293–304, Oct. 2022.

[39] H. Fang, P. Xiao, L. Xu, and Y. Chen, “Clip2video: Mastering video-text retrieval via image CLIP,” 2021, arXiv:2106.11097.

[40] Y. Ma, G. Xu, X. Sun, M. Yan, J. Zhang, and R. Ji, “X-CLIP: End-to-end multi-grained contrastive learning for video-text retrieval,” in Proc. ACM Int. Conf. Multimedia (MM), 2022, pp. 1–13.

[41] M. Tang, Z. Wang, Z. Liu, F. Rao, D. Li, and X. Li, “CLIP4Caption: CLIP for video caption,” in Proc. ACM Int. Conf. Multimedia (MM), 2021, pp. 4858–4862.

[42] Z. Wang et al., “CRIS: CLIP-driven referring image segmentation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2022, pp. 11676–11685.

[43] B. Ni et al., “Expanding language-image pretrained models for general video recognition,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2022, pp. 1–18.

[44] R. Zhang et al., “PointCLIP: Point cloud understanding by CLIP,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2022, pp. 8542–8552.

[45] Y. Rao et al., “DenseCLIP: Language-guided dense prediction with context-aware prompting,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2022, pp. 18061–18070.

[46] L. Zhang, X. Wang, D. V. Kalashnikov, S. Mehrotra, and D. Ramanan, “Query-driven approach to face clustering and tagging,” IEEE Trans. Image Process., vol. 25, no. 10, pp. 4504–4513, Oct. 2016.

[47] H. Tang, C. Yuan, Z. Li, and J. Tang, “Learning-attention-guided pyramidal features for few-shot fine-grained recognition,” Pattern Recognit., vol. 130, Oct. 2022, Art. no. 108792.
**Neng Dong** is currently pursuing the Ph.D. degree with the School of Computer Science and Engineering, Nanjing University of Science and Technology. His current research interests include computer vision and deep learning, especially object re-identification and adversarial learning.

**Liyan Zhang** received the Ph.D. degree in computer science from the University of California at Irvine, Irvine, CA, USA, in 2014. She is currently a Professor with the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, China. Her current research interests include multimedia analysis, computer vision, and deep learning. She received the Best Paper Award from ICMR in 2013 and the Best Student Paper Award from MMM in 2016.

**Jinhui Tang** (Senior Member, IEEE) received the B.E. and Ph.D. degrees from the University of Science and Technology of China, Hefei, China, in 2003 and 2008, respectively. He is currently a Professor with the Nanjing University of Science and Technology, Nanjing, China. He has authored more than 200 papers in top-tier journals and conferences. His current research interests include multimedia analysis and computer vision. He is a fellow of IAPR. He was a recipient of the Best Paper Award in the 2007 ACM MM and the 2020 ACM MM Asia and the Best Paper Runner-Up in the 2015 ACM MM. He has served as an Associate Editor for *IEEE Transactions on Neural Networks and Learning Systems*, *IEEE Transactions on Knowledge and Data Engineering*, *IEEE Transactions on Multimedia*, and *IEEE Transactions on Circuits and Systems for Video Technology*. 