Interactive Sketch-based Person Re-Identification with Text Feedback

Xinyi Wu[1,2], Cuiqun Chen[3], Hui Zeng[1], Zhiping Cai[1]*, Bo Du[2] and Mang Ye[2]*
[1]College of Computer Science and Technology, National University of Defense Technology, Changsha, China
[2]School of Computer Science, Wuhan University, Wuhan, China
[3]School of Computer Science and Technology, Anhui University, Hefei, China
{wuxinyi17, zenghui116, zpcai}@nudt.edu.cn, chencuiqun@ahu.edu.cn, {dubo, yemang}@whu.edu.cn

***Abstract*—Sketch-based Person Re-identification (Sketch ReID) aims to retrieve a person of interest across disjoint cameras using hand-drawn sketches as queries. A significant issue is the limited structural clues of sketch queries, which fail to convey high-level semantic retrieval intentions, such as colors and genders. Existing works typically combine sketches and texts for multi-modal retrieval, which inevitably introduces modality interference and relies heavily on expensive tri-modal datasets. In this paper, we propose, for the first time, an interactive and flexible sketch-based person retrieval framework that incorporates user feedback to refine the sketch person retrieval ranking without text training. A lightweight vision-to-text converting network is proposed to represent sketches with equivalent pseudo-word tokens, which aims to provide context assistance for interactive retrieval. Then, the sketch token can be seamlessly integrated with text feedback tokens within CLIP's textual space for explicit sketch-text compositionality, thus achieving feedback-guided ranking refinement. Extensive experiments underscore the superiority of our InteractReID. Code will be available at https://github.com/littlexinyi/InteractReID.**

***Index Terms*—Sketch-based person ReID, Interactive retrieval**

I. Introduction

Person Re-identification (ReID) [1] involves identifying target persons from massive cross-camera videos using query clues, which has wide applications in intelligent video surveillance, criminal investigation, and other fields. Traditional image-to-image ReID methods [2]–[4] use photos of targets as query clues. Despite significant progress made, a crucial issue that was often overlooked lies with the availability of such photo queries — the target's visual photos are often not readily accessible. Therefore, Sketch ReID [5] was introduced to match candidate pedestrian photos using sketch queries, where sketches play a role as user-provided clues to capture intricate visual details with similar structures.

However, as Figure 1 (a) shows, although sketch and image modalities are both visual expressions, the sketch-based query lacks high-level semantic information such as color, posture, and gender. The bottleneck is that a single sketch query can not fully express the user's all retrieval intentions and capture false positive samples with similar structure appearance, thus resulting in limited retrieval performance.

*Corresponding authors.

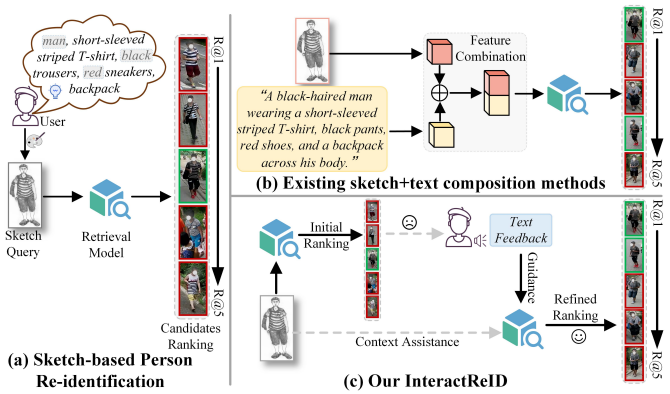

Fig. 1. (a) Sketch query is inherently inadequate in representing high-level semantic features (gender, color), which will erroneously retrieve some false positive persons with similar structural appearance (clothing type, length). (b) Existing works conduct simple combination between sketch and text for retrieval enhancement, which may introduce feature interference due to the significant modality gap. (c) Our InteractReID achieves flexible interactive person retrieval through context assistance and feedback guidance, enabling sketch and text to query in synergy within the textual domain.

Given that natural language is a direct medium for users to express their intentions, existing works [6], [7] have introduced natural language to alleviate the above issue, where they combined the two descriptive modalities as a powerful query through feature concatenation for multi-modal joint retrieval in a supervised manner (Figure 1 (b)). However, the simple combination may distort the optimal sketch-text composed semantics due to modality interference. Meanwhile, the supervised training paradigm requires labor-intensive tri-modal datasets, making the overall framework less flexible.

In this paper, we for the first time propose InteractReID, *an interactive and flexible sketch person retrieval framework* that can integrate the user's text feedback to refine the sketch person retrieval ranking without text training. Figure 1 (c) shows a detailed interaction process. Merely using a sketch for person retrieval may not produce ideal ranking results due to the missing semantic intents. In our framework, users can provide text feedback with high-level semantic details based on the initial retrieval results. This allows the model to refine the ranking and interactively improve retrieval performance.

Specifically, we achieve interactive and flexible retrieval through Context Assistance and Feedback Guidance. Retrieval

ranking refinement is conducted on CLIP's excellent image-text-aligned embedding space [8], where a textual feedback-guided query is provided to retrieve and rank gallery images with higher feature similarities. Meanwhile, to simultaneously utilize the sketch information for retrieval, inspired by the implicit grammatical composition capability of CLIP's text encoder [8], [9], our idea is to represent the sketch as a pseudo-word token to provide context assistance for interactive retrieval. The sketch token can be seamlessly and flexibly integrated with user-guided text feedback through textual token concatenation, enabling sketch and text to query in synergy within the image-text-aligned embedding space.

The above explicit sketch-text compositionality can ingeniously avoid modality interference and also achieve optimal composed semantic mining with CLIP's textual domain. Meanwhile, a vision-to-text converting network is trained to identify semantic-equivalent tokens for sketch modalities. The total training process only involves self-alignment of the sketch and retrieval alignment between sketches and images, eliminating the need for extensive textual descriptions.

The main contributions can be summarized as follows:

- We introduce a novel interactive person retrieval framework for sketch ReID, flexibly integrating the user's text feedback with sketch queries for ranking refinement.
- InteractReID enables interactive retrieval through context assistance and feedback guidance, which not only achieves optimal semantic composition but also reduces the training reliance on text modality.
- Extensive quantitative and qualitative results from both sketch-based and interactive person retrieval scenarios highlight the superiority of InteractReID.

II. Related Work

A. Sketch-based Person Re-identification

Sketch ReID was first introduced in [5] along with the proposal of the PKUSketch dataset. Most works [10]–[12] are devoted to alleviating the modality gap for feature alignment, where a joint embedding space is built through adversarial feature learning, semantic consistency building, or auxiliary modality generation. Lin *et al.* [13] focuses on the sketch style subjectivity problem with the proposal of the multi-style Market-Sketch-1K dataset. Zhai *et al.* [6] explores the complementary semantics of sketch and text, and conducts descriptive query fusion for higher retrieval accuracy. Following this trend, Chen *et al.* [7] proposes a unified person ReID framework with descriptive queries, which can effectively handle varying multi-modal data.

However, existing sketch+text composition methods focus on modality late-fusion (*i.e.*, simple feature combination), which will introduce modality interference and thus distort the optimal sketch-text composed semantics. However, our InteractReID aims to mine the optimal composed semantics through modality early-fusion (*i.e.*, representing sketch as a pseudo-word token and conducting token concatenation with the text modality), achieving interactive and flexible sketch person retrieval with text feedback.

B. Interactive Image Retrieval with user feedback

Interactive image retrieval seeks to refine the ranking results by incorporating user feedback with their intentions, which is popular in fashion search and recommendation fields. Existing works focused on learning with feedback in various forms, such as sketches [14], spatial layouts [15], attributes [16], or texts [17]–[21]. Among them, natural language is widely utilized. A classic approach is to design an image-text composition module that generates composed queries similar to the target image. VAL [18] proposes multiple composition modules to conduct fine-grained fusion. Additionally, CosMo [19] considers this task more comprehensively from the image's style and content aspects and proposes a content-style modulator for detailed combination. FashionVLP [20] firstly proposes a VLP transformer-based model, conducting composed retrieval with the help of prior knowledge from large corpora. FashionNTM [22] extends the interaction to multi-turn retrieval via cascaded memories.

However, these works are mainly limited to the fashion retrieval field and conduct modality late-fusion. Our InteractReID firstly introduces the idea of feedback refinement into the sketch person retrieval field. Different from the above innovations, we represent sketches as pseudo-word tokens to achieve sketch-text synergy during interactive inference, where the sketch tokens can be seamlessly integrated with user-provided text feedback tokens with the help of CLIP's implicit grammatical composition capability.

III. Methodology

A. Revisiting CLIP

CLIP [8] is a typical dual-stream vision-language pre-trained model with separate modality encoders, trained on 400 million image-text pairs [8] with a multi-class N-pair contrastive loss to learn a modality-aligned joint embedding space. In the training process, the cosine similarity of N paired image and text embeddings is maximized, while the cosine similarity of other $N^2 - N$ unpaired embeddings is minimized. The image encoder $\mathbf{V}$ empirically adopts Vision Transformer (ViT) [23], which firstly conducts patch embedding for an input image I and then passes them into the vision transformer to obtain its visual feature $i^v = \mathbf{V}(I) \in \mathbb{R}^d$. The text encoder $\mathbf{T}$ conducts similar processing on input word sequences W. Words in the sequence will be tokenized by byte pair encoding (BPE) and then embedded as a high-dimensional vector for further feature learning into Transformer [24], which produces final textual feature $w^t = \mathbf{T}(W) \in \mathbb{R}^d$.

B. Task-oriented Knowledge Adaptation for CLIP

CLIP [8] has demonstrated powerful cross-modal semantic alignment capabilities, achieving remarkable performance across a wide range of downstream cross-modal tasks [25], [26]. However, it is unfeasible to directly apply CLIP to downstream multi-modal ReID tasks due to the significant domain gap. Therefore, we first conduct task-oriented knowledge adaptation for CLIP through parameter fine-tuning on the large-scale sketch-image-text tri-modal dataset Tri-PEDES [7],

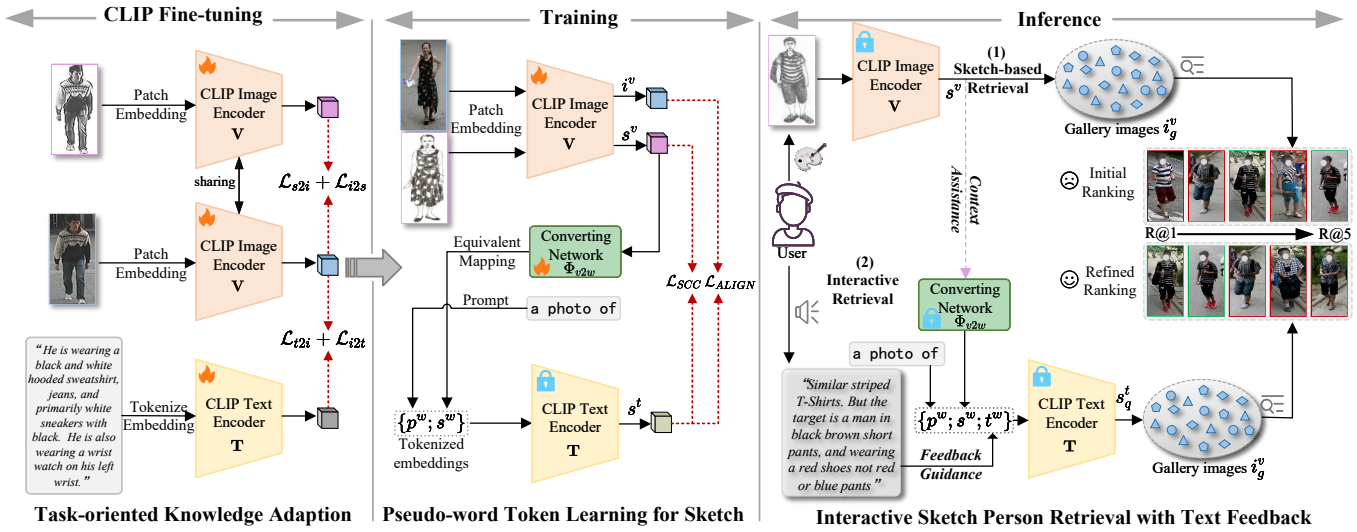

Fig. 2. Training and inference of our proposed InteractReID framework. **Left**: We first fine-tune CLIP on the multi-modal ReID task for downstream knowledge adaptation, where the cross-modal matching loss $\mathcal{L}_{CM} = \mathcal{L}_{t2i} + \mathcal{L}_{i2t} + \mathcal{L}_{s2i} + \mathcal{L}_{i2s}$ is used for modality alignment. Based on the knowledge-adapted CLIP, we aim to find the sketch's equivalent mapping in textual space for interactive retrieval, where two contrastive losses $\mathcal{L}_{SCC}$ and $\mathcal{L}_{ALIGN}$ are utilized for sketch's pseudo-word token generation. **Right**: During inference, basic sketch person retrieval is conducted through similarity calculation. To interactively refine the retrieval ranking, users' text feedback t^w can be integrated with the sketch's pseudo-word token s^w through context and feedback guidance.

aiming to further utilize CLIP's modality-aligned knowledge in the downstream multi-modal ReID tasks.

As Figure 2 shows, given a batch of B sketch-image-text pairs, we equip with a cross-modal matching loss to pull positive pairs together and push negative pairs apart. Specifically, different modalities are associated by embedding their cosine similarity distributions into the KL divergence.

Using text-to-image matching as an example, for each pair $\langle w_i^t, i_j^v \rangle$, we model its matching probability through the feature's cosine similarity, which can be calculated by:

$$p_{i,j} = \frac{\exp\left(sim\left(w_i^t, i_j^v\right)/\tau\right)}{\sum_{k=1}^{B} \exp\left(sim\left(w_i^t, i_k^v\right)/\tau\right)}, \tag{1}$$

$$sim\left(w_i^t, i_j^v\right) = \frac{\left(w_i^t\right)^\top i_j^v}{\|w_i^t\| \left\|i_j^v\right\|}, \tag{2}$$

where τ is a temperature parameter that controls the probability distribution peaks. Then the matching loss from text to image can be computed by:

$$\mathcal{L}_{t2i} = D_{\mathrm{KL}}\left(\hat{q}_{i,j} \| p_{i,j}\right) = -\frac{1}{|B|} \sum_{i=1}^{B} \sum_{j=1}^{B} \hat{q}_{i,j} \log p_{i,j}, \tag{3}$$

where $\hat{q}_{i,j}$ is the true matching probability, which is the normalized ground-truth labels $q_{i,j} / \sum_{k=1}^{B} q_{i,k}$. ($q_{i,j} = 1$ means a matched pair with the same identity, and $q_{i,j} = 0$ indicates the unmatched pair).

Symmetrically, the matching loss from image to text $\mathcal{L}_{i2t}$, from sketch to image $\mathcal{L}_{s2i}$, and from image to sketch $\mathcal{L}_{i2s}$ can also be calculated by modeling matching probability between modality pairs. The total cross-modal matching loss for fine-tuning CLIP to align sketch-image-text can be formulated as:

$$\min_{\{\mathbf{V},\mathbf{T}\}} \quad \mathcal{L}_{CM} = \mathcal{L}_{t2i} + \mathcal{L}_{i2t} + \mathcal{L}_{s2i} + \mathcal{L}_{i2s}. \tag{4}$$

C. Pseudo-word Token Learning for Sketch

To achieve interactive sketch person retrieval with user's text feedback, based on the vision-text joint embedding space provided by CLIP, we aim to find a pseudo-word token that can accurately capture the sketch semantics for context assistance, thus achieving explicit sketch-text compositionality through flexible textual token concatenation.

Concretely, we train a lightweight vision-to-text converting network $\mathbf{\Phi_{v2w}}$ with 1-layer MLP on downstream realistic sketch retrieval datasets to achieve the equivalent mapping. Given an input sketch query S, we first obtain its visual embedding through CLIP's vision encoder: $s^v = \mathbf{V}(S) \in \mathbb{R}^d$, which will be sent to the converting network $\mathbf{\Phi_{v2w}}$ to generate its equivalent pseudo-word token embedding as $s^w = \mathbf{\Phi_{v2w}}(s^v) \in \mathbb{R}^d$. To maintain the semantic integrity and compositionality, inspired by the popular prompt learning paradigm [27], we attach s^w at the end of a universal prompt sentence p^w (*e.g.*, "a photo of", "an image of") and pass it through CLIP's text encoder to obtain the sketch's final language-equivalent feature $s^t = \mathbf{T}(\{p^w; s^w\}) \in \mathbb{R}^d$. Revisiting our training objective, we aim to bring the sketch features in the textual space s^t as close as possible to those in the visual space s^v. To achieve this, we propose a self-cycle contrastive loss to impose training constraints on the converting network $\mathbf{\Phi_{v2w}}$, *i.e.*,

$$\min_{\{\mathbf{\Phi},\mathbf{V}\}} \quad \mathcal{L}_{SCC} = \mathcal{L}_{cst}(s^v, s^t) + \mathcal{L}_{cst}(s^t, s^v), \tag{5}$$

$$\mathcal{L}_{cst}(s^v, s^t) = -\frac{1}{|B|} \sum_{i \in B} \log \frac{\exp\left((s_i^v)^\top s_i^t/\tau\right)}{\sum_{j \in B} \exp\left((s_i^v)^\top s_j^t/\tau\right)}, \tag{6}$$

$$\mathcal{L}_{cst}(s^t, s^v) = -\frac{1}{|B|} \sum_{i \in B} \log \frac{\exp\left((s_i^t)^\top s_i^v/\tau\right)}{\sum_{j \in B} \exp\left((s_i^t)^\top s_j^v/\tau\right)}. \tag{7}$$

TABLE I
PERFORMANCE COMPARISONS WITH SOTA METHODS ON MARKET-SKETCH-1K. † MEANS RE-IMPLEMENTATION BY US.

Methods	Retrieval Type	R@1	R@5	R@10	mAP	mINP
DCLNet [28]	Sketch-based	12.24	29.20	39.58	13.45	-
DSCNet [29]		13.84	30.55	40.34	14.73	-
DEEN [30]		12.11	25.44	30.94	12.62	-
BDG [13]		18.10	38.95	50.75	19.61	-
UniReID† [7]		8.65	19.45	26.88	11.16	7.52
UniReID† [7]	Sketch+Text	27.01	53.97	67.68	31.48	23.33
InteractReID	Sketch-based	37.30	61.81	73.80	39.65	30.49
	Interactive	**50.42**	**79.70**	**88.61**	**54.98**	**46.88**

Meanwhile, the learned token embedding s^t for sketch should also be aligned with its paired images $i^v = \mathbf{V}(I) \in \mathbb{R}^d$ through the following feature alignment loss:

$$\min_{\{\mathbf{\Phi},\mathbf{V}\}} \quad \mathcal{L}_{ALIGN} = \mathcal{L}_{cst}(s^t, i^v) + \mathcal{L}_{cst}(i^v, s^t). \tag{8}$$

We update the parameters of the proposed vision-to-text converting network $\mathbf{\Phi_{v2w}}$ and CLIP image encoder $\mathbf{V}$ on downstream sketch person retrieval datasets Market-Sketch-1K [13] and PKUSketch [5] without paired textual descriptions, while keeping the CLIP text encoder $\mathbf{T}$ frozen to fully utilize its grammatical composition capability for sketch and text's token concatenation.

D. Inference

The retrieval practicality of our proposed InteractReID framework can be evaluated from not only sketch-based person retrieval but also interactive person retrieval.

Firstly, the sketch query's feature s^v is extracted and compared with gallery features i_g^v for similarity calculation, thus generating the initial retrieval ranking results, which are then returned to users for retrieval accuracy evaluation.

In order to refine the retrieval ranking results, users can provide any textual feedback t^w for interactive communication. Specifically, we introduce context assistance through the well-trained vision-to-text converting network $\mathbf{\Phi_{v2w}}$ for sketch's equivalent pseudo-word token generation: $s^w = \mathbf{\Phi_{v2w}}(s^v)$. Then the feedback guidance is achieved through textual token concatenation among p^w, s^w, and t^w. Therefore, the composed query feature can be generated through the CLIP text encoder: $s_q^t = \mathbf{T}(\{p^w; s^w; t^w\}) \in \mathbb{R}^d$. Lastly, we conduct refined retrieval ranking by comparing the cosine similarity between s_q^t and gallery visual features i_g^v.

IV. EXPERIMENTS

A. Experimental Settings

1) Datasets and evaluation metrics: General R@K, mAP, and mINP are utilized for performance evaluation. Higher R@K, mAP, and mINP indicate better retrieval performance.

Tri-PEDES [7] is a mixture of three multi-modal ReID datasets, where sketches are generated according to RGB images in CUHK-PEDES, ICFG-PEDES, and RSTPReid. Tri-PEDES is utilized for CLIP fine-tuning and consists of 21203 identities, 115233 RGB images, and sketches with 175972 text descriptions.

TABLE II
PERFORMANCE COMPARISONS WITH SOTA METHODS ON PKUSKETCH.

Methods	Retrieval Type	R@1	R@5	R@10	mAP	mINP
AFLNet [5]	Sketch-based	34.00	56.30	72.50	-	-
LMDI [31]		49.00	70.40	80.20	-	-
CDAC [10]		60.80	80.60	88.80	-	-
UniReID [7]		69.80	88.60	95.80	72.97	68.25
BDG [13]		70.00	-	-	66.37	-
CCSC [11]		86.00	98.00	100.00	83.70	-
DALNet [12]		90.00	98.60	100.00	86.20	-
TriReID [6]	Sketch+Text	20.00	50.00	64.00	-	-
UniReID [7]		91.40	**98.80**	99.80	**91.76**	88.97
InteractReID	Sketch-based	89.00	98.00	99.00	88.90	85.11
	Interactive	**92.00**	98.60	**99.80**	91.14	**89.69**

PKUSketch [5] is a basic dataset for sketch ReID, which totally contains 200 identities, each of which has two RGB images and a sketch. 150 persons are randomly selected for training, and the remaining 50 persons are used for testing. We conduct 10 experiments on randomly partitioned datasets and take the average for evaluation.

Market-Sketch-1K [13] is a new sketch ReID dataset with 6 unique painting styles for each identity, which consists of 4,763 sketches of 996 identities and 32,668 photos of 1,501 identities in total. Sketches are divided into 6 groups according to different painting styles, where each style group contains 498 disjointed identities for training and testing.

2) Implementation Details: We utilize CLIP (ViT-B/16) [8] for feature extraction and grammatical composition. We first conduct task-oriented knowledge adaptation on Tri-PEDES [7], which aims to fine-tune CLIP with AdamW with 1e-4 for merely 5 epochs on 4 RTX 4090 24GB GPUs with a batch size of 240. All input images are resized to 224 × 224 and augmented through random horizontal flipping. The maximum length of the textual token sequence is 77. τ is set to 0.07. In the pseudo-word token learning process, 1-layer MLP $\mathbf{\Phi_{v2w}}$ with 768 hidden units and ReLU activation is trained on Market-Sketch-1K or PKUSketch with 20 epochs for realistic scenario simulation. We use AdamW with 1e-5 for $\mathbf{V}$ and 1e-4 for $\mathbf{\Phi_{v2w}}$ training. During inference, textual annotations for both datasets labeled by [6], [32] are used to simulate users' text feedback. Note that our InteractReID does not involve any text annotations while training $\mathbf{\Phi_{v2w}}$.

B. Comparison with State-of-the-art Methods

We compare the proposed InteractReID with existing reproducible baselines on Market-Sketch-1K and PKUSketch datasets, focusing on two realistic retrieval scenarios using different queries: sketch-based and interactive retrieval. As shown in Table I and Table II, our InteractReID has demonstrated its superiority in both of the above scenarios.

Concretely, when first using the sketch-based query for person retrieval, it can be observed that our paradigm outperforms other baselines significantly on both datasets, indicating its excellent modality alignment capability. The performance gain is likely due to the effectiveness of our cross-modal matching and feature alignment losses.

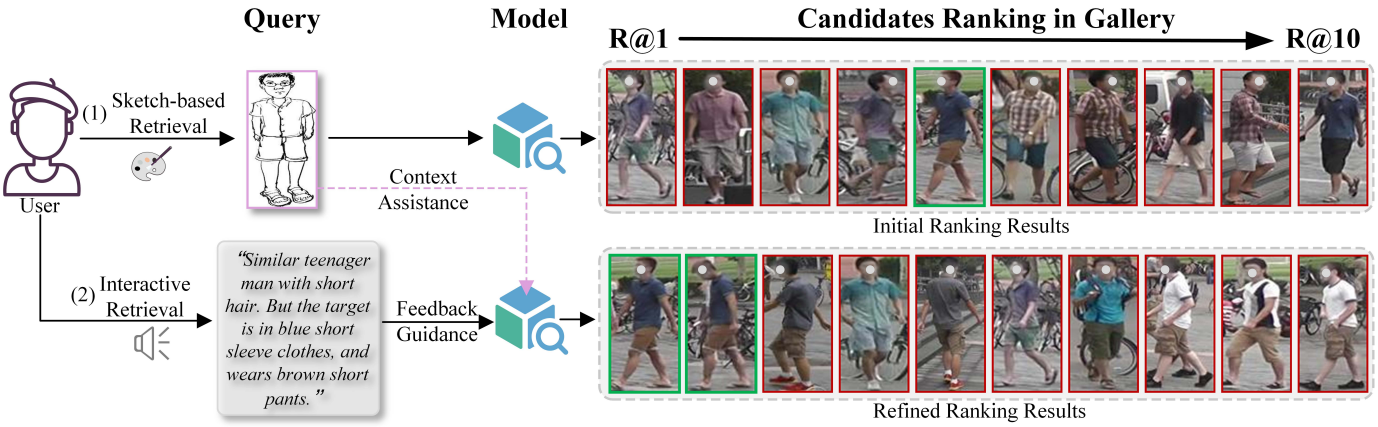

Fig. 3. A typical interactive person retrieval process of our InteractReID. Top-10 initial and refined ranking results on the Market-Sketch-1K dataset are compared through sketch-based and interactive person retrieval with text feedback. Gallery samples in green boxes match the query, while in red boxes mismatch the query. Best viewed in color and zoomed in.

TABLE III
ABLATIONS ON THE NECESSITY AND DATASETS OF TASK-ORIENTED KNOWLEDGE ADAPTATION.

No.	Adaptation Datasets	Sketch-based			Interactive		
		R@1	mAP	mINP	R@1	mAP	mINP
1	-	22.87	26.80	18.96	25.44	30.45	22.62
2	Tri-CUHK-PEDES	34.09	36.08	25.98	44.56	48.65	39.61
3	Tri-ICFG-PEDES	31.77	34.53	25.54	30.08	35.53	27.44
4	Tri-RSTPReid	30.04	33.09	23.82	31.81	36.34	27.77
5	Tri-PEDES	**37.30**	**39.65**	**30.49**	**50.42**	**54.98**	**46.88**

Meanwhile, our InteractReID can also achieve interactive person retrieval through context assistance and text feedback guidance. Compared to existing sketch+text composition methods which adopt modality late-fusion for retrieval (*i.e.*, simple feature combination), InteractReID adopts modality early-fusion that achieves explicit sketch-text compositionality through textual token concatenation. Thanks to the interaction capability achieved by integrating sketch's pseudo-word tokens with user-provided text feedback in CLIP's language token space, InteractReID surpasses other supervised composition methods with a R@1 of 92.00% on PKUSketch and 49.24% on Market-Sketch-1K for retrieval ranking refinement.

C. Ablation Study

We comprehensively evaluate the effectiveness of the proposed InteractReID on Market-Sketch-1K from the following three questions.

1) Is Task-oriented Knowledge Adaptation necessary? As shown in Table III, we first eliminate the knowledge adaptation fine-tuning process and directly apply CLIP's pre-trained parameters for downstream converting network training (No.1). This leads to significant performance degradation compared to No.5 in both the sketch-based and interactive retrieval scenarios, which highlight the importance of task-oriented knowledge adaptation. Furthermore, we conduct ablation studies on the adaptation datasets (No.2 - No.5 in Table III), where Tri-PEDES is a combination of Tri-CUHK-PEDES, Tri-ICFG-PEDES, and Tri-RSTPReid. Experimental results demonstrate that the complexity of adaptation datasets is strongly associated with the model's generalization ability in downstream retrieval tasks.

TABLE IV
ABLATIONS ON THE DESIGN OF VISION-TO-TEXT CONVERTING NETWORK. l DENOTES THE LAYER OF MLP BLOCKS, AND h MEANS THE HIDDEN UNITS OF MLP. THE UNIT FOR "TRAIN PARAMS" IS **'M'**.

No.	Network Design	Train Params	Sketch-based			Interactive		
			R@1	mAP	mINP	R@1	mAP	mINP
1	-	86.19	28.86	30.92	21.06	33.84	39.26	31.63
2	Linear only	86.46	34.77	37.72	28.40	45.95	50.97	42.48
3	l=1, h=512	86.72	36.29	39.20	27.74	49.24	53.06	44.85
4	l=1, h=768	86.98	37.30	39.65	30.49	**50.42**	**54.98**	**46.88**
5	l=1, h=1024	87.24	38.31	40.38	30.88	46.79	51.74	44.08
6	l=2, h=512	86.98	33.97	36.91	28.13	44.43	50.12	42.14
7	l=2, h=768	87.57	37.76	40.11	30.94	46.71	51.76	43.97
8	l=2, h=1024	88.29	**38.95**	**40.90**	**31.29**	46.62	50.54	42.08
9	l=3, h=512	87.24	29.75	33.88	25.72	44.35	49.17	40.82
10	l=3, h=768	88.16	35.19	37.91	28.94	48.06	52.74	44.21
11	l=3, h=1024	89.34	34.60	38.37	29.54	45.49	50.11	41.78

2) Is Vision-to-text Converting Network effective? The converting network is designed to achieve feature-equivalent mapping from visual to textual space. To verify the above motivation, we conduct ablation studies in Table IV from the following three designs: directly sending visual features into the text encoder without network converting (No.1), mapping with a linear-only layer without activation (No.2), different MLP designs for the converting network (No.3 - No.11).

There are three observations: (1) Learnable parameters are necessary to map visual features into the textual embedding space due to the significant modality gap. (2) Non-linear ReLU activation is critical for enhancing the expressing capability of the converting network. (3) Considering the balance between retrieval performance and training efficiency, 1-layer MLP with 768 hidden units (No.4) is selected as the optimal network configuration.

We also experimented with other variants, such as varying the dropout rate, but did not observe any significant improve-

ments.

3) Does pseudo-word tokens really capture sketch semantics? To analyze the effectiveness of the pseudo-word token in capturing sketch information, we evaluate the well-trained model using relative validation datasets. Concretely, we utilize sketches' pseudo-word tokens as queries to retrieve in the gallery which solely consists of input sketches. R@1 of 95.64% and R@5 of 99.76% on Market-Sketch-1K demonstrate that the semantic effectiveness of pseudo-word tokens.

D. Qualitative Results

A typical interactive person retrieval process is demonstrated in Figure 3 to verify the practicality of our InteractReID. Firstly, basic sketch-based person retrieval is conducted, which tends to match persons only with similar structures. When the initial ranking results are returned, our framework can support users to provide discriminative semantic feedback for interactive and flexible retrieval. Concretely, when mapping the sketch into its equivalent pseudo-word token and combining it with feedback tokens in the textual domain, InteractReID can achieve retrieval ranking refinement.

V. Conclusion and Future works

In this paper, we propose a novel interactive sketch-based person retrieval framework InteractReID, which can flexibly integrate sketch queries with user's text feedback to achieve retrieval ranking refinement. Concretely, Task-oriented Knowledge Adaptation is first conducted for CLIP's pre-trained knowledge transfer. Based on this, a vision-to-text converting network is trained to represent sketch as a pseudo-word token so that the CLIP text encoder can flexibly compose the sketch token and text feedback tokens for interactive retrieval. Extensive quantitative and qualitative results on both sketch-based and interactive person retrieval scenarios demonstrate our superiority. Comprehensive ablation studies are conducted to validate the effectiveness of the proposed InteractReID framework. In the future, we aim to explore LLM-based conversational interaction, allowing for multi-turn retrieval based on user's text feedback.

Acknowledgment

This work was supported by the National Natural Science Foundation of China under Grants (62172155, 62472434, 62102425, 62176188, 62361166629, 62225113, 62306215), and the Science and Technology Innovation Program of Hunan Province (2022RC3061 and 2023RC3027).

References

[1] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi, "Deep learning for person re-identification: A survey and outlook," *IEEE TPAMI*, vol. 44, no. 6, pp. 2872–2893, 2021.

[2] Hao Ni, Jingkuan Song, Xiaopeng Luo, Feng Zheng, Wen Li, and Heng Tao Shen, "Meta distribution alignment for generalizable person re-identification," in *Proc. of CVPR*, 2022, pp. 2487–2496.

[3] Siyuan Li, Li Sun, and Qingli Li, "Clip-reid: exploiting vision-language model for image re-identification without concrete text labels," in *Proc. of AAAI*, 2023, vol. 37, pp. 1405–1413.

[4] Haidong Zhu, Pranav Budhwant, Zhaoheng Zheng, and Ram Nevatia, "Seas: Shape-aligned supervision for person re-identification," in *Proc. of CVPR*, 2024, pp. 164–174.

[5] Lu Pang, Yaowei Wang, Yi-Zhe Song, Tiejun Huang, and Yonghong Tian, "Cross-domain adversarial feature learning for sketch re-identification," in *Proc. of ACM MM*, 2018, pp. 609–617.

[6] Yajing Zhai et al., "Trireid: Towards multi-modal person re-identification via descriptive fusion model," in *Proc. of ICMR*, 2022, pp. 63–71.

[7] Cuiqun Chen et al., "Towards modality-agnostic person re-identification with descriptive query," in *Proc. of CVPR*, 2023, pp. 15128–15137.

[8] Alec Radford, Jong Wook Kim, et al., "Learning transferable visual models from natural language supervision," in *Proc. of ICML*. PMLR, 2021, pp. 8748–8763.

[9] Alberto Baldrati et al., "Zero-shot composed image retrieval with textual inversion," in *Proc. of ICCV*, 2023, pp. 15338–15347.

[10] Fengyao Zhu, Yu Zhu, et al., "Cross-domain attention and center loss for sketch re-identification," *IEEE TIFS*, vol. 17, pp. 3421–3432, 2022.

[11] Yafei Zhang, Yongzeng Wang, et al., "Cross-compatible embedding and semantic consistent feature construction for sketch re-identification," in *Proc. of ACM MM*, 2022, pp. 3347–3355.

[12] Xingyu Liu, Xu Cheng, et al., "Differentiable auxiliary learning for sketch re-identification," in *Proc. of AAAI*, 2024, pp. 3747–3755.

[13] Kejun Lin, Zhixiang Wang, Zheng Wang, Yinqiang Zheng, and Shin'ichi Satoh, "Beyond domain gap: Exploiting subjectivity in sketch-based person retrieval," in *Proc. of ACM MM*, 2023, pp. 2078–2089.

[14] Subhadeep Koley, Ayan Kumar Bhunia, et al., "You'll never walk alone: A sketch and text duet for fine-grained image retrieval," in *Proc. of CVPR*, 2024, pp. 16509–16519.

[15] Arko Barman and Shishir K Shah, "A graph-based approach for making consensus-based decisions in image search and person re-identification," *IEEE TPAMI*, vol. 43, no. 3, pp. 753–765, 2019.

[16] Bo Zhao, Jiashi Feng, Xiao Wu, and Shuicheng Yan, "Memory-augmented attribute manipulation networks for interactive fashion search," in *Proc. of CVPR*, 2017, pp. 1520–1528.

[17] Xiaoxiao Guo, Hui Wu, et al., "Dialog-based interactive image retrieval," in *Proc. of NeurIPS*, 2018.

[18] Yanbei Chen et al., "Image search with text feedback by visiolinguistic attention learning," in *Proc. of CVPR*, 2020, pp. 3001–3011.

[19] Seungmin Lee, Dongwan Kim, and Bohyung Han, "Cosmo: Content-style modulation for image retrieval with text feedback," in *Proc. of CVPR*, 2021, pp. 802–812.

[20] Sonam Goenka, Zhaoheng Zheng, et al., "Fashionvlp: Vision language transformer for fashion retrieval with feedback," in *Proc. of CVPR*, 2022, pp. 14105–14115.

[21] Dafeng Li and Yingying Zhu, "Visual-linguistic alignment and composition for image retrieval with text feedback," in *Proc. of ICME*. IEEE, 2023, pp. 108–113.

[22] Anwesan Pal et al., "Fashionntm: Multi-turn fashion image retrieval via cascaded memory," in *Proc. of ICCV*, 2023, pp. 11323–11334.

[23] Alexey Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[24] A Vaswani, "Attention is all you need," *Proc. of NeurIPS*, 2017.

[25] Ding Jiang and Mang Ye, "Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval," in *Proc. of CVPR*, 2023, pp. 2787–2797.

[26] Xinyi Wu, Wentao Ma, et al., "Text-based occluded person re-identification via multi-granularity contrastive consistency learning," in *Proc. of AAAI*, 2024, pp. 6162–6170.

[27] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu, "Conditional prompt learning for vision-language models," in *Proc. of CVPR*, 2022, pp. 16816–16825.

[28] Hanzhe Sun, Jun Liu, et al., "Not all pixels are matched: Dense contrastive learning for cross-modality person re-identification," in *Proc. of ACM MM*, 2022, pp. 5333–5341.

[29] Yiyuan Zhang et al., "Dual-semantic consistency learning for visible-infrared person re-identification," *IEEE TIFS*, vol. 18, pp. 1554–1565, 2022.

[30] Yukang Zhang and Hanzi Wang, "Diverse embedding expansion network and low-light cross-modality benchmark for visible-infrared person re-identification," in *Proc. of CVPR*, 2023, pp. 2153–2162.

[31] Shaojun Gui et al., "Learning multi-level domain invariant features for sketch re-identification," *Neurocomputing*, vol. 403, pp. 294–303, 2020.

[32] Yutian Lin et al., "Improving person re-identification by attribute and identity learning," *Pattern Recognition*, vol. 95, pp. 151–161, 2019.