Optimal Transport-based Labor-free Text Prompt Modeling for Sketch Re-identification

Rui Li,[*] **Tingting Ren,**[*] **Jie Wen,**[†] **Jinxing Li**[†]
School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen
lrhit@stu.hit.edu.cn, {rentt0410,lijinxing158}@gmail.com, jiewen_pr@126.com

Abstract

Sketch Re-identification (Sketch Re-ID), which aims to retrieve target person from an image gallery based on a sketch query, is crucial for criminal investigation, law enforcement, and missing person searches. Existing methods aim to alleviate the modality gap by employing semantic metrics constraints or auxiliary modal guidance. However, they incur expensive labor costs and inevitably omit fine-grained modality-consistent information due to the abstraction of sketches. To address this issue, this paper proposes a novel *Optimal Transport-based Labor-free Text Prompt Modeling* (OLTM) network, which hierarchically extracts coarse- and fine-grained similarity representations guided by textual semantic information without any additional annotations. Specifically, multiple target attributes are flexibly obtained by a pre-trained visual question answering (VQA) model. Subsequently, a text prompt reasoning module employs learnable prompt strategy and optimal transport algorithm to extract discriminative global and local text representations, which serve as a bridge for hierarchical and multi-granularity modal alignment between sketch and image modalities. Additionally, instead of measuring the similarity of two samples by only computing their distance, a novel triplet assignment loss is further proposed, in which the whole data distribution also contributes to optimizing the inter/intra-class distances. Extensive experiments conducted on two public benchmarks consistently demonstrate the robustness and superiority of our OLTM over state-of-the-art methods.

1 Introduction

With the growing need for urban public safety, traditional person re-identification (Re-ID) methods [1, 2, 3] are gradually becoming inadequate for criminal investigations and missing person tracking, as the individuals of interest may not have been captured by surveillance cameras. To bolster social security management and combat criminal activities, sketch person re-identification (Sketch Re-ID), which utilizes eyewitness clues to draw professional sketches as queries and match target images in a photo gallery database, has received widespread attention from researchers and scholars [4, 5, 6], as shown in Fig. 1. Nonetheless, due to the considerable disparity in modal heterogeneity resulting from the varied sketch styles of different artists and the

Figure 1: The illustration of sketch Re-ID. Different artists create sketches based on clues provided by witness to assist the police in identifying targets.

---

[*]Equal Contribution.

[†]Corresponding Author.

diverse postures of real pedestrians in monitoring, sketch Re-ID remains highly challenging and necessitates further exploration and investigation.

Giving the high generalization and inherent abstraction of person characteristics in sketches, the pedestrian features depicted in a single sketch (such as clothing and gender) may match multiple similar real images, as illustrated in Fig. 1. A viable solution is to leverage the inter-modality interaction within the feature space to achieve hard alignment. Such methods typically employ loss constraints to directly map different modalities into a generic latent space [4, 7, 8, 9]. However, this hard alignment manner may not fully capture the complex dependencies and correlations that exist within and across modalities. To compensate for the lack of details in the above manner, another branch is to introduce an intermediate modality to bridge two source modalities. For instance, [10, 11] generate simulated sketches through adversarial learning, but the generated sketches are inevitably corrupted with noise due to limited generation performance. Additionally, [7, 12] construct benchmarks that contain textual information to alleviate modal gap; [6] improves inference efficiency by introducing text only during the training process. Despite the fact that these additional texts do contribute to mitigating the modal gap, they are all manually labeled, requiring significant human labor in real-world applications. Moreover, existing text-guided methods [6, 7] only focus on global text embeddings as masks, neglecting finer and richer local features. Therefore, this paper aims to address two key challenges: **i) developing sufficient textual information as a transition mechanism without incurring additional costs**, and **ii) further exploring fine-grained discriminative information for multi-granularity interaction.**

To address the above issues, we propose the Optimal Transport-based Labor-free Text Prompt Modeling (OLTM) framework, which implicitly incorporates text semantic information during training, facilitating hierarchical and multi-granularity modal alignment. In particular, OLTM is composed of three main components: i) text prompt reasoning (TPR); ii) text-injected coarse-grained alignment module (TCA); iii) consensus-guided fine-grained interaction module (CFI). On the one hand, to introduce text sequences without additional manual annotation, we dynamically transfer pre-trained language-visual knowledge into the downstream task. Specifically, TPR first generates multi-dimensional person attributes based on real images with a pre-trained visual question answering (VQA) model. Then, these attributes are inserted as fixed parts into learnable prompts to obtain the textual embedding representations. TCA integrates global parts of the embeddings to achieve the coarse-grained alignment across modalities. On the other hand, to explore fine-grained information, we employ optimal transport theory to enhance deep-level interaction. Concretely, TPR formulates the mapping from local parts of the textual embeddings to more discriminative feature representations, i.e., consensus, as an optimal transport problem. Subsequently, guided by consensus, CFI selectively focuses on key details, extracting fine-grained conceptual representations for sketch-ID. In addition, due to the significant heterogeneity gap between modalities, using Euclidean distance as a sole metric to measure feature similarity is inadequate. Thus, we propose a triplet assignment loss to optimize feature distance measurement and improve model performance. Extensive experiments are conducted on two challenging benchmarks, which demonstrate the favorable comparison of OLTM with other state-of-the-art methods.

The core contributions of this work are summarized as follows: (1) This paper proposes a novel optimal transport-based labor-free text prompt modeling framework for sketch Re-ID. To our best knowledge, this is *the first attempt* to apply VQA-generated text responses as a means to achieve modal alignment in sketch Re-ID *without any additional annotations*. (2) A novel text prompt reasoning module is deployed to dynamically extract global textual embeddings and discriminative fine-grained consensus, which guide the hierarchical multi-granularity alignment module in injecting semantic knowledge into the modeling process. (3) A new triplet assignment loss is proposed, which optimizes inter-/intra-class distance by considering overall data distribution information.

2 Related Work

Sketch Re-identification As an important part of public safety guarantee, sketch Re-ID is a novel and challenging task that aims to match a person image with given professional sketches. Existing sketch Re-ID methods could be roughly classified into two groups according to their interaction modes, i.e., hard alignment methods [4, 8, 9] and soft alignment methods [12, 10, 6]. The former try to learn modality embeddings in a common latent space by employing some modality interaction operations or semantic metrics. Pang *et al*. [4] pioneered a sketch-photo benchmark and introduced

cross-domain adversarial learning to narrow the feature gap. Zhang *et al.* [9] proposed an advanced cross-modal learning mechanism for handling non-corresponding information between modalities. However, due to significant differences between modalities, this direct alignment paradigm inevitably loses fine-grained modality-specific cues [13]. Hence, some of the latter methods investigate gentler alignment techniques through transitional modality. For example, Chen *et al.* [10] designed a dynamic updatable auxiliary sketch modality to increases the diversity of training samples; Zhai *et al.* [12] introduced a multi-modal Re-ID task by combining text and sketch as query for retrieval, exploiting their complementary advantages. Obviously, auxiliary modalities lacking detailed information may introduce noise, while data annotation is a labor-intensive task. In this paper, we *first attempt* to use text attributes generated by a reasonable VQA model as guidance for achieving multi-granularity alignment across modalities in sketch Re-ID.

Optimal Transport For optimizing the moving cost between distributions, Optimal Transport (OT) was first proposed by Kantorovich [14], which has shown significant potential in machine learning and computer vision, e.g., domain adaptation [15, 16, 17], learning with noisy labels [18, 19], and feature matching [20, 21]. Zhang *et al.* [22] incorporated OT into the re-ranking phase of image retrieval, significantly improving accuracy and efficiency. Similarly, Sergio *et al.* [23] first applied OT in visual place recognition and introduced a novel local feature aggregation method. In semi-supervised person Re-ID, OT often achieves the mapping between pseudo labels and classes as a classifier [24, 25]. In addition, Ling *et al.* [26] designed a assignment strategy for alleviating the intra-identity variations; Wasserstein distance was used to rectify the original global distance between samples and provides aligned distance estimation for local features [27]. Considering the enormous potential of OT in feature aggregation and distribution mapping, our study adopts OT to assist fine-grained alignment between modalities and guide the model in extracting the overall sample distribution pattern.

Prompt Learning Prompt learning initially garnered widespread attention and extensive research in natural language processing [28, 29, 30], which has gradually demonstrated significant potential in vision-language (V-L) models [31, 32, 33] and pure vision models [34, 35, 36]. Prompt learning provides a flexible way to adapt pre-trained models to downstream tasks by training only additional parameters. This enables prompts to capture task-specific information while guiding the fixed model's performance [37, 38]. In sketch-based image retrieval, [39] innovatively learns a unified prompt for different branches in CLIP's [40] visual encoding layer, fully exploiting CLIP's zero-shot learning potential. In text-to-image person Re-ID, [41] introduces a multi-prompt strategy to integrate text prompts from various sources for fine-grained interaction. Furthermore, Li et al. [42] first attempt to conduct in-depth research on zero-shot multi-modal ReID through a large foundational model. In this paper, we delve into the significant application of prompt learning in sketch Re-ID, innovatively generating global text representations by integrating fixed and learnable prompts, and utilizing OT to reason consensus for effectively guiding detailed interactions across modalities.

3 Preliminaries

3.1 Problem Statement

To ensure clarity, we represent the gallery containing m images I as $\mathcal{G} = \{I_i, y_i\}_{i=1}^m$ and the query set containing n sketches S as $\mathcal{S} = \{S_j, y_j\}_{j=1}^n$, where $y \in \{1, \ldots, C\}$ are the identity labels for C distinct pedestrian entities. Notably, each entity may include multiple images and sketches. The goal of Sketch Re-ID is to retrieve pedestrian images from the gallery $\mathcal{G}$ that match one or multiple given sketch. Like [6], there exist two types of query methods: single sketch query and multiple sketches query. This section will use single sketch query as an example, and the same applies to multiple sketches query. Formally, we define a matching function $\mathcal{M} : \mathcal{S} \times \mathcal{G} \to \mathbb{R}^{n \times m}$ that assigns a similarity score to each pair (I_i, S_j). The objective is to learn a function $\mathcal{M}$ such that for any sketch S_j and image I_i,

$$\mathcal{M}(I_i, S_j) > \mathcal{M}(I_k, S_j) \quad if \quad y_i = y_j \ and \ y_k \neq y_j, \tag{1}$$

where I_i and I_k are images from the gallery set, and y_i and y_k are their respective identity labels.

3.2 Optimal Transport

Optimal Transport is a mathematical theory that focuses on finding an efficient solution between two probability distributions, minimizing the cost of transporting one distribution into another. We briefly review the theoretical derivation of optimal transport. Let $\Gamma_r := \{\boldsymbol{x} \in \mathbb{R}_+^r | \boldsymbol{x}^\top \mathbf{1}_r = 1\}$ represents
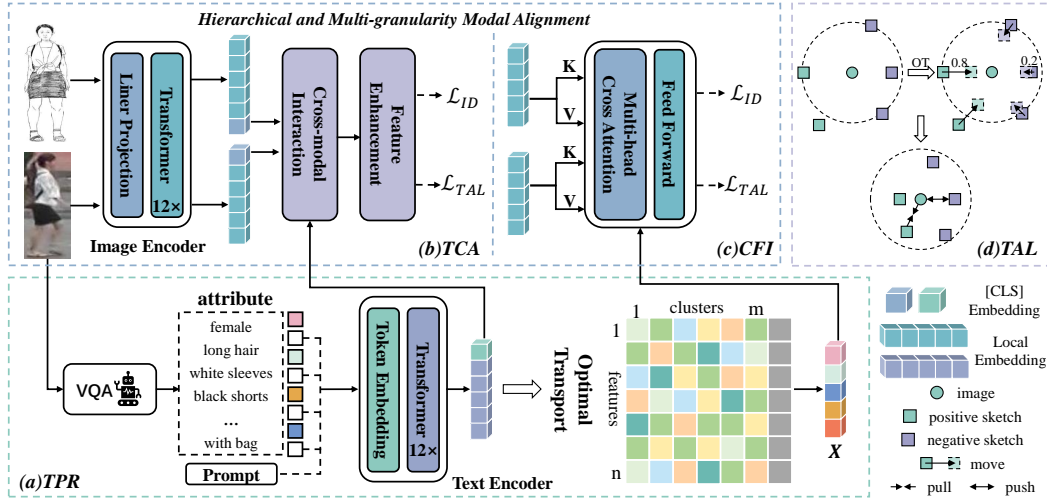
Figure 2: Overview of our proposed OLTM network. Our model includes four main parts, i.e., text prompt reasoning (TPR), text-injected coarse-grained alignment Module (TCA), consensus-guided fine-grained interaction module (CFI) and triplet assignment loss (TAL). Specifically, TPR flexibly generates target characteristics through VQA, and combines prompt learning and optimal transport to reason text global embedding and local consensus. TCA and CFI extract modality-specific representations from image and sketch modalities to achieve hierarchical and multi-granularity alignment. Finally, TAL is designed to optimize distance measurement between samples and improve the model's capacity to capture local relationships.

the probability simplex, where $\mathbf{1}_r$ is the r-dimensional vector of ones. Given two probability simplex vectors $\boldsymbol{\alpha} \in \Gamma_m$ and $\boldsymbol{\beta} \in \Gamma_n$ and a cost matrix $\boldsymbol{C} \in \mathbb{R}^{m \times n}$, the objective of OT is to seek the optimal transport plan $\boldsymbol{P}^*$ mapping $\boldsymbol{\alpha}$ to $\boldsymbol{\beta}$ at the minimum cost:

$$d_{\boldsymbol{C}}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \min_{\boldsymbol{P} \in U(\boldsymbol{\alpha}, \boldsymbol{\beta})} \langle \boldsymbol{C}, \boldsymbol{P} \rangle,$$
$$U(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \left\{ \boldsymbol{P} \in \mathbb{R}_+^{m \times n} \mid \boldsymbol{P}\mathbf{1}_n = \boldsymbol{\alpha}, \boldsymbol{P}^\top \mathbf{1}_m = \boldsymbol{\beta} \right\}, \tag{2}$$

where $U(\boldsymbol{\alpha}, \boldsymbol{\beta})$ denotes the transport polytope of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, i.e., the solution space of $\boldsymbol{P}$. The above problem is to find optimal solution $\boldsymbol{P}^*$ in a set of all possible joint probabilities of (X, Y), where X and Y represent random variables with marginal distribution $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$.

Eq. 8 indicates that OT is a linear programming problem which is theoretically solvable in polynomial time, but its complexity becomes prohibitively high as the feature dimension increases [43]. To this end, Sinkhorn algorithm [44] adopts an iterative strategy to obtain the optimal solution $\boldsymbol{P}^* = Diag(\boldsymbol{u})\boldsymbol{K}Diag(\boldsymbol{v})$ with near-square complexity [45]. $\boldsymbol{u}$ and $\boldsymbol{v}$ can be solved through alternately iterating the following two equations: $\boldsymbol{u}^{(z)} = \boldsymbol{\alpha}/(\boldsymbol{K}\boldsymbol{v}^{(z-1)})$ and $\boldsymbol{v}^{(z)} = \boldsymbol{\beta}/(\boldsymbol{K}^\top \boldsymbol{u}^{(z)})$, where $\boldsymbol{K} = exp(\boldsymbol{C}/\epsilon)$, ϵ is the regularization coefficient and z is the iterations (cf. **Appendix**). Since this method integrates the importance of all features when solving the optimal solution, it can analyze the overall data distribution.

4 The Proposed Method

4.1 Overall Architecture

Fig. 2 provides an overview of the OLTM architecture. The image and text encoders discussed in this paper are based on CLIP [40], and any language-visual model utilizing a Transformer architecture may also be employed. Notably, the image encoders utilized for both images and sketches employ shared weights to ensure the mapping of features into a unified semantic space. For an input RGB image, we obtain embeddings $\boldsymbol{R} = \{\boldsymbol{R}_{\text{cls}}, \boldsymbol{r}_1, \ldots, \boldsymbol{r}_p\} \in \mathbb{R}^{(p+1) \times d}$ through the image encoder, where p is the number of non-overlapping patches, $\boldsymbol{R}_{\text{cls}}$ and $\boldsymbol{R}_{\text{local}} = \{\boldsymbol{r}_i\}_{i=1}^p$ represent d-dimensional global and local features, respectively. Similarly, the embeddings of a sketch can be represented as $\boldsymbol{S} = \{\boldsymbol{S}_{cls}, \boldsymbol{S}_{local}\} = \{\boldsymbol{S}_{\text{cls}}, \boldsymbol{s}_1, \ldots, \boldsymbol{s}_p\} \in \mathbb{R}^{(p+1) \times d}$. Firstly, to provide reasonable text semantic

guidance, Text Prompt Reasoning (TPR) generates attribute descriptions about pedestrians based on the RGB image, and obtains the textual embeddings $\boldsymbol{T} = \{\boldsymbol{T}_{eos}, \boldsymbol{T}_{local}\}$ through prompt learning. Then, TPR extracts the fine-grained consensus $\boldsymbol{X}$ from $\boldsymbol{T}_{local}$ through clustering. Subsequently, Text-injected Coarse-grained Alignment (TCA) module embeds global contextual information $\boldsymbol{T}_{eos}$ into visual features $\boldsymbol{R}_{\text{cls}}$ and $\boldsymbol{S}_{\text{cls}}$. Meanwhile, Consensus-guided Fine-grained Interaction (CFI) module utilizes $\boldsymbol{X}$ to address fine-grained semantic misalignment between $\boldsymbol{R}_{local}$ and $\boldsymbol{S}_{local}$. Additionally, due to the significant differences between sketches and RGB images, Euclidean distance between independent samples may ignore the influence of overall sample distribution. Thus, we introduce a more comprehensive distance measurement method and propose triplet assignment loss $\mathcal{L}_{tal}$. During training, all these modules will be jointly optimized through identity loss [46] $\mathcal{L}_{id}$ and $\mathcal{L}_{tal}$: $\mathcal{L}_{OLTM} = \mathcal{L}_{id} + \eta \mathcal{L}_{tal}$, where η is a scaling factor. During inference, only $\boldsymbol{R}_{\text{cls}}$ and $\boldsymbol{S}_{\text{cls}}$ are used to match queries for practical application requirements.

4.2 Text Prompt Reasoning

Significant image differences and inherent abstract nature, cause semantic misalignment during knowledge acquisition, severely impacting model's reasoning and generalization capabilities. To address this issue, TPR introduces intermediate modality to guide alignment between modalities without additional costs. Moreover, TPR employs a dynamic consensus acquisition strategy to enhance the discriminative power of local text features.

Text Attribute Generation Sketches, unlike conventional Re-ID tasks, are vulnerable to subjective emotions and drawing skills of artists, leading to a lack of detailed information crucial for model learning. The text's objectivity and flexibility prompt the model to focus more on semantic contextual information during knowledge acquisition. However, directly generating a comprehensive textual description of pedestrian images inevitably introduces irrelevant noise, thereby reducing model performance. Therefore, we retain the advanced modeling capabilities of large-scale language-visual models for images as possible. Specifically, for a given RGB image, TPR utilizes a pre-trained visual question answering model to address k specific details (cf. **Appendix**) and acquire corresponding descriptions for the target: $att = \{att_1, att_2, \ldots, att_k\}$. Importantly, this process introduces textual detail guidance during model training but excludes text-related components during inference.

Learnable Prompt Strategy Inspired by [47], we combine the learnable prompt with original text attributes, without incurring additional expert knowledge compared to the handcrafted prompt. Concretely, TPR initially transform these attributes into tokens through CLIP tokenizer, i.e., $\boldsymbol{a} = Tokenizer(att)$. Then, l learnable prompts $\{p_1, p_2, \ldots, p_l\}$ are embedded into these fixed attributes tokens, forming the textual description: $\boldsymbol{q} = \{p_1, a_1, p_2, a_2, \ldots, p_l, a_k\}$. This integration introduces a dynamic knowledge learning mechanism that reduces noise introduction compared to handcrafted prompts, while enhancing the flexibility of modal interaction and transferability of text embeddings. Subsequently, the whole token $\boldsymbol{q}$ is fed into a frozen text encoder to generate text embeddings $\boldsymbol{T} = \{\boldsymbol{T}_{\text{sos}}, \boldsymbol{t}_1, \boldsymbol{t}_2, \ldots, \boldsymbol{t}_n, \boldsymbol{T}_{\text{eos}}\} \in \mathbb{R}^{(n+2) \times d}$, where $\boldsymbol{T}_{\text{sos}}$ and $\boldsymbol{T}_{\text{eos}}$ denote the [SOS] and [EOS] token, n is the number of d-dimensional word tokens. Based on widely-used token selection, $\boldsymbol{T}_{\text{eos}}$ serves as the global feature, while $\boldsymbol{T}_{\text{local}} = \{\boldsymbol{t}_j\}_{j=1}^{n}$ represents a sequence of basic local tokens.

Dynamic Consensus Acquisition To more effectively address fine-grained semantic variations (e.g., hats, shoes) across modalities, we explore methods to filter out non-informative features for enhancing the representational capacity of text embedding for detailed information. Therefore, based on local textual feature $\boldsymbol{T}_{local}$, we employ metric learning to draft a dynamic consensus acquisition strategy for capturing the discriminative prototypical representations $\boldsymbol{X}$.

To begin, in order to adaptively learn related-task knowledge, local text representations $\boldsymbol{T}_{local} \in \mathbb{R}^{n \times d}$ are fed into consensus multi-layer perceptron (ConMLP) blocks to achieve feature enhancement. Then, a prototypical descriptor is formed by assigning all enhanced features to a set of atoms. The cost matrix $\boldsymbol{C} \in \mathbb{R}_{+}^{n \times m}$ can be calculated for assignment, where the (i, j)-th element $\boldsymbol{C}_{i,j}$ represents the cost of assigning a feature to an atom. In other words, $\boldsymbol{C}$ evaluates the affinity between the enhanced features and the prototypical atoms. Concretely, to introduce certain priors as guidance [23], the enhanced features are used to learn the cost matrix through two fully connected layers initialized randomly. In addition, some irrelevant information in textual features, such as those representing association, may disrupt the model's ability to learn the target details. Inspired by the solutions used in graph matching and key-point matching [20, 23], we set a learnable "bin" in $\boldsymbol{C}$. Due to the differences between prior distributions, non-informative features can be assigned to it. Specifically,

we extend the cost matrix from $\boldsymbol{C}$ to $\bar{\boldsymbol{C}} = [\boldsymbol{C}, \bar{\boldsymbol{c}}] \in \mathbb{R}_+^{n \times (m+1)}$, and $\bar{\boldsymbol{c}} = w\mathbf{1}_n$, where w is a learnable parameter and $\mathbf{1}_n = [1, \ldots, 1]^\top \in \mathbb{R}^n$ represents n-dimensional vector of ones. Following Sec. 3.2, we consider that assignment where the enhanced features' mass, $\boldsymbol{\alpha} = \mathbf{1}_n$, should be assigned to the atoms or the "bin", $\boldsymbol{\beta} = [\mathbf{1}_m^\top, n - m]^\top$, is an optimal transport problem:

$$d_{\bar{\boldsymbol{C}}}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \min_{\boldsymbol{P} \in \boldsymbol{U}(\boldsymbol{\alpha}, \boldsymbol{\beta})} \langle \bar{\boldsymbol{C}}, \boldsymbol{P} \rangle,$$
$$\boldsymbol{U}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \{\boldsymbol{P} \in \mathbb{R}_+^{n \times (m+1)} | \boldsymbol{P}\mathbf{1}_{m+1} = \boldsymbol{\alpha}, \boldsymbol{P}^\top \mathbf{1}_n = \boldsymbol{\beta}\}. \tag{3}$$

The optimal assignment plan can be solved by Sinkhorn algorithm (cf. **Appendix**) through iterative strategy. For better descriptors quality, the extra "bin" is discarded to obtain the optimal assignment $\boldsymbol{P} \in \mathbb{R}_+^{n \times m}$. Finally, the augmented comprehensive representation, i.e., consensus $\boldsymbol{X} \in \mathbb{R}^{m \times d}$, is obtained by aggregating enhanced textual features and optimal transport plan: $\boldsymbol{X} = \boldsymbol{P}^\top \boldsymbol{T}_{local}$.

4.3 Hierarchical and Multi-granularity Modal Alignment

TPR develops a robust transitional means to facilitate reasonable and efficient modal alignment. In order to leverage effectively the multi-granularity textual information, TCA utilizes global embedding to achieve coarse-grained modal interaction, while CFI optimizes fine-grained alignment based on enhanced comprehensive representation.

Text-injected Coarse-grained Alignment Module TCA utilizes transformer architecture, known for its efficiency in modeling long-distance dependencies, to capture global information, as shown in Fig. 2. Furthermore, we implement a cross-attention mechanism using global text embedding to emphasize contextual information injection at the beginning. For an input pair $(\boldsymbol{R}, \boldsymbol{S})$, cross-attention mechanism computes the *query* using text global representation $\boldsymbol{T}_{eos}$ and derives the *key* and *value* from modality-specific global features $\boldsymbol{R}_{cls}$ and $\boldsymbol{S}_{cls}$:

$$\boldsymbol{Q}_{R/S} = \boldsymbol{T}_{eos} \cdot \boldsymbol{w}^Q, \boldsymbol{K}_{R/S} = (\boldsymbol{R}/\boldsymbol{S})_{cls} \cdot \boldsymbol{w}^K, \boldsymbol{V}_{R/S} = (\boldsymbol{R}/\boldsymbol{S})_{cls} \cdot \boldsymbol{w}^V,$$
$$CA(\boldsymbol{R}/\boldsymbol{S}, \boldsymbol{T}_{eos}) = Attention(\boldsymbol{Q}_{R/S}, \boldsymbol{K}_{R/S}, \boldsymbol{V}_{R/S}), \tag{4}$$

where $\boldsymbol{R}/\boldsymbol{S}$ signifies identical operations across both modalities; $\boldsymbol{w}^Q$, $\boldsymbol{w}^K$ and $\boldsymbol{w}^V$ denote shared learnable parameters, while $\boldsymbol{Q}_{R/S}$, $\boldsymbol{K}_{R/S}$ and $\boldsymbol{V}_{R/S}$ represent *query*, *key* and *value* for either the RGB or sketch modality, respectively. After obtaining the fusing features that contain textual knowledge, TCA introduces the standard transformer blocks to refine modality-specific features. Finally, we can acquire the final global concept representations $\boldsymbol{R}'_{cls}$ and $\boldsymbol{S}'_{cls}$ for sketch Re-ID.

Consensus-guided Fine-grained Interaction Module Due to the inherent complexity of sketches and RGB images and the potential for semantic misalignment during learning, capturing detail variations in modalities is crucial. Fortunately, the fine-grained consensus $\boldsymbol{X}$ provided by TPR, which contains rich detail information, offers a key solution to this problem. CFI adopts a transformer structure based on multi-head cross attention, and it converts the original local feature $\boldsymbol{R}_{local}$ and $\boldsymbol{S}_{local}$ to more discriminative representations through for robust Re-ID:

$$\hat{\boldsymbol{Q}}_{R/S} = \boldsymbol{X} \cdot \boldsymbol{w}^{\hat{Q}}, \hat{\boldsymbol{K}}_{R/S} = (\boldsymbol{R}/\boldsymbol{S})_{local} \cdot \boldsymbol{w}^{\hat{K}}, \hat{\boldsymbol{V}}_{R/S} = (\boldsymbol{R}/\boldsymbol{S})_{local} \cdot \boldsymbol{w}^{\hat{V}},$$
$$Head_h^{R/S} = Attention(\hat{\boldsymbol{Q}}_{R/S}, \hat{\boldsymbol{K}}_{R/S}, \hat{\boldsymbol{V}}_{R/S}), \tag{5}$$
$$MH(\boldsymbol{R}/\boldsymbol{S}, \boldsymbol{X}) = Concat(Head_1^{R/S}, \ldots, Head_H^{R/S}),$$

where $\boldsymbol{w}^{\hat{Q}}$, $\boldsymbol{w}^{\hat{K}}$ and $\boldsymbol{w}^{\hat{V}}$ denote the parameters of project layers of h-th head for both modalities. As a result, the text detail information about pedestrian characteristics can assist our model to address the problem caused by the diversity and uncertainty of sketches and RGB images. Eventually, the local concept representations $\boldsymbol{R}'_{local}$ and $\boldsymbol{S}'_{local}$ can be obtained.

4.4 Triplet Assignment Loss

The triple loss, a commonly-used matching loss in cross-modal learning, performs well in performance through adjusting the distance of hardest negatives in scenarios like image-text matching [12, 48] and video-text retrieval [49]. However, this strategy independently trains each semantic part with equal contribution, disregarding the overall data distribution impact when multiple samples from

Table 1: Comparison with state-of-the-art methods on Market-Sketch-1K dataset. Both training and testing set uses all sketches. 'S' and 'M' represent single-query and multi-query, respectively. 'Backbone' refers to network structure used by each method, mainly including: ResNet50 [50] and CLIP [40]. **Bold** values represent the optimal results.

Methods	Query	Backbone	Reference	mAP	Rank@1	Rank@5	Rank@10
DDAG [51]	S	ResNet50	ECCV'2020	12.13	11.22	25.40	35.02
CM-NAS [52]	S	ResNet50	ICCV'2021	0.82	0.70	2.00	3.90
CAJ [53]	S	ResNet50	ICCV'2021	2.38	1.48	3.97	7.34
MMN [54]	S	ResNet50	MM'2021	10.41	9.32	21.98	29.58
DART [55]	S	ResNet50	CVPR'2022	7.77	6.58	16.75	23.42
DCLNet [56]	S	ResNet50	MM'2022	13.45	12.24	29.20	39.5
DSCNet [57]	S	ResNet50	TIFS'2022	14.73	13.84	30.55	40.34
DEEN [58]	S	ResNet50	CVPR'2023	12.62	12.11	25.44	30.94
BDG [6]	S	ResNet50	MM'2023	19.61	18.10	38.95	50.75
	M			24.45	24.70	50.40	63.45
UNIReID [7]	S	CLIP	CVPR'2023	34.97	31.52	57.17	70.46
	M			55.18	56.63	82.33	91.97
OLTM (Ours)	S	CLIP	-	**38.35**	**36.75**	**63.88**	**74.05**
	M			**62.55**	**69.48**	**90.36**	**95.18**

different modalities exhibit slight differences. This oversight may lead to inaccurate estimation of sample distances and potentially result in sub-optimal local minima [27]. To this end, we propose a new triplet assignment loss (TAL) to establish a more rational measure for evaluating the proximity of local features.

For an input positive pair $(\boldsymbol{R}_i, \boldsymbol{S}_i)$ in a mini-batch x, the feature representations obtained through model inference are $\boldsymbol{R}_i^{'}$ and $\boldsymbol{S}_i^{'}$. If we treat the feature sets of all samples for each of two modalities in x as two discrete distributions, their alignment can be considered an optimal transport problem. The cost matrix $\hat{\boldsymbol{C}}$ is derived from pairwise feature similarities: $\hat{\boldsymbol{C}}_{i,j} = [(\boldsymbol{R}_i^{'})^{\top} \boldsymbol{S}_j^{'}]_+$. We aim to acquire the optimal transport matrix $\boldsymbol{P}^*$ with the least amount of cost, where $\boldsymbol{P}_{i,j}^*$ represents the assignment weight of $(\boldsymbol{R}_i, \boldsymbol{S}_j)$ obtained after balancing the overall distribution. TAL can be represented based on triplet loss as the weighted sum of the original distance and the optimal assignment distance, dynamically updated at a certain rate γ:

$$\mathcal{L}_{tal}(\boldsymbol{R}_i, \boldsymbol{S}_i) = [m - D(\boldsymbol{R}_i, \boldsymbol{S}_i) + D(\boldsymbol{R}_i, \hat{\boldsymbol{S}}_h)]_+ + [m - D(\boldsymbol{R}_i, \boldsymbol{S}_i) + D(\hat{\boldsymbol{R}}_h, \boldsymbol{S}_i)]_+,$$
$$D(\boldsymbol{R}_i, \boldsymbol{S}_i) = \gamma E(\boldsymbol{R}_i, \boldsymbol{S}_i) + (1 - \gamma)(1 - \boldsymbol{P}_{i,i}^*)E(\boldsymbol{R}_i, \boldsymbol{S}_i) \qquad (6)$$

where $[x]_+ = \max(x, 0)$, $\hat{\boldsymbol{R}}_h = argmax_{R_j \neq R_i} D(\boldsymbol{R}_j, \boldsymbol{S}_i)$ and $\hat{\boldsymbol{S}}_h = argmax_{S_j \neq S_i} D(\boldsymbol{R}_i, \boldsymbol{S}_j)$ are the most similar negatives in x for $(\boldsymbol{R}_i, \boldsymbol{S}_i)$, and $E(\boldsymbol{R}_i, \boldsymbol{S}_i) = \|\boldsymbol{R}_i^{'} - \boldsymbol{S}_i^{'}\|_2$ denotes the Euclidean distance between feature representations.

5 Experiment

5.1 Experiment Setup

Datasets Two publicly available benchmark datasets, namely PKU-Sketch [4] and Market-Sketch-1K [6], are utilized for performance evaluation. Both of them are sketched and annotated by professional artists. **PKU-Sketch** is the first publicly Sketch Re-ID dataset, containing data for 200 pedestrians, with each individual being represented through one sketch and two photos. In accordance with the setting of [4], we randomly select 150 identities for training and 50 for testing, and final results are derived from the average of 10 experimental runs. **Market-Sketch-1K** is a large-scale dataset derived from the Market-1501 [1], which is created by six artists based on descriptions, featuring multiple perspectives and artistic styles. The training set consists of 2,332 sketches and 12,936 photos, while the testing set comprises 2,375 sketches and 19,732 photos. Following the experimental setup in [6], our method will be evaluated in three settings: single-query, multi-query, and cross-style retrieval.

Table 2: Comparison with state-of-the-art methods on PKU-Sketch dataset. 'Backbone' includes GoogleNet [62], VGG-16 [63], ResNet50, ViT [64], and CLIP. '-' denotes the unavailable results. '†' indicates that we reproduce UNIReID results following our training configuration.

Methods	Backbone	Reference	mAP	Rank@1	Rank@5	Rank@10
TripleSN [65]	-	CVPR'2016	-	9.0	26.8	42.2
GNSiamese [66]	GoogleNet	TOG'2016	-	28.9	54.0	62.4
AFLNet [4]	GoogleNet	MM'2018	-	34.0	56.3	72.5
LMDI [8]	VGG-16	Neuro'2020	-	49.0	70.4	80.2
SKetchTrans [10]	ViT	MM'2022	-	84.6	94.8	98.2
CCSC [9]	ViT	MM'2022	83.7	86.0	98.0	**100.0**
SKetchTrans+ [5]	ViT	PAMI'2023	-	85.8	96.0	99.0
UNIReID† [7]	CLIP	CVPR'2023	88.7	92.4	98.0	99.6
DALNet [11]	ResNet50	AAAI'2024	86.2	90.0	98.6	**100.0**
OLTM (Ours)	CLIP	-	**91.4**	**94.0**	**99.4**	**100.0**

(a) Market-Sketch-1K (Single-Query)   **(b) Market-Sketch-1K (Multi-Query)**   **(c) PKU-Sketch**
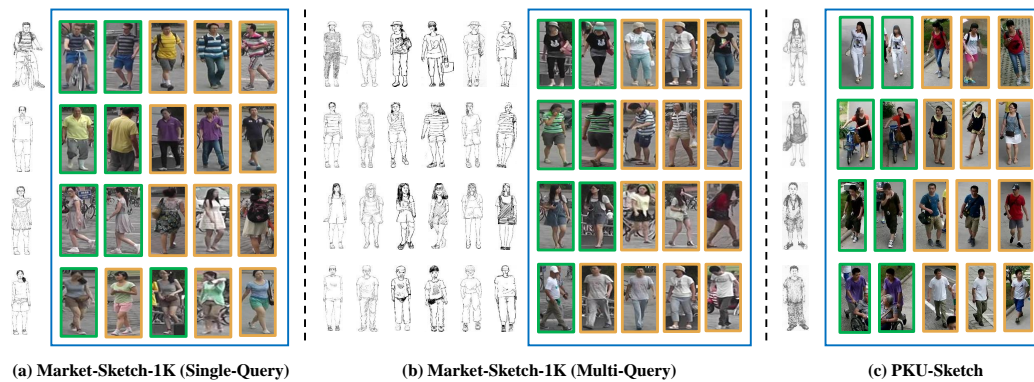
Figure 3: The Rank-5 retrieval results on two datasets. For the Market-Sketch-1K dataset, both single-query and multi-query scenarios are presented. **Green** border indicates correctly retrieved target pedestrians, while **yellow** border indicates incorrectly matched pedestrians.

Evaluation Metrics In line with [4, 9, 6, 59], we use Rank-k metrics ($k = 1, 5, 10$) and mean Average Precision (mAP) as evaluation metrics. The higher values of the above three metrics, the better performance.

Implementation Details OLTM uses a pre-trained CLIP-ViT-B/16 [40] as image encoder, and extracts text features with a pre-trained CLIP Text Transformer. Fine-grained text attributes are derived from a ViLT-based [60] VQA model. Importantly, our VQA model is replaceable. To avoid the cost overhead from multiple calls to VQA model during training and inference, text attribute generation is performed during the data processing stage. For multi-query scenarios, we input a weighted sum of multiple sketches. Input images are resized to 288×144, and augmented with random horizontal flipping and style augmentation [61]. More experimental configuration details are available in the supplementary materials.

5.2 Comparison with State-of-the-Art Methods

Performance on Market-Sketch-1K. The experimental results for Market-Sketch-1K are shown in Tab. 1. OLTM significantly outperforms all state-of-the-art methods in both single-query and multi-query settings. In the single-query scenario, OLTM achieves an mAP of 38.35% and a Rank-1 of 36.75%, surpassing state-of-the-art method by 3.38% and 5.23%, respectively. In the multi-query scenario, OLTM achieves an mAP of 62.55% and a Rank-1 of 69.48%, exceeding state-of-the-art methods by 7.37% and 12.85%, respectively. This result is primarily due to the introduction of textual semantic information in the training process of OLTM, which implicitly guides images and sketches to focus on modality-invariant features. Through hierarchical and multi-granularity alignment, the model is able to uncover discriminative fine-grained information, leading to more

Table 3: Ablation studies on Market-Sketch-1K dataset. Training and testing are under the multi-query setting. "Handcrafted" and "VQA" denote manually annotated and VQA generated text attributes, respectively. "Template" represents the sentence template defined by experts. "Prompt" denotes the learnable text prompts. The 'Baseline' uses an image encoder to process both modalities and employs simple cross-attention to integrate the global features. '$\mathcal{L}_{\text{htl}}$' [67] represents the hard triplet loss. **Bold** values represent the optimal results.

Prompt setting				Module			Loss			Metrics	
Handcrafted	VQA	Template	Prompt	Baseline	TCA	CFI	$\mathcal{L}_{\text{ID}}$	$\mathcal{L}_{\text{htl}}$	$\mathcal{L}_{\text{tal}}$	mAP	Rank@1
-	-	-	-							55.47	60.04
✓	-	✓	-	✓	✓	✓	✓	-	✓	61.46	68.07
✓	-	-	✓							61.81	67.47
-	✓	✓	-							61.76	65.46
-	✓	-	✓	✓	-	-	✓	-	✓	57.74	60.84
				✓	✓	-				61.10	65.66
-	✓	-	✓	✓	✓	✓	✓	-	-	54.93	57.83
				✓	✓	✓	✓	✓	-	61.63	66.06
-	✓	-	✓	✓	✓	✓	✓	-	✓	**62.55**	**69.48**

accurate queries. The visualization results on Market-Sketch-1K are shown in Fig. 3. When sketch possesses clearly distinguishable features, query results are satisfactory. In contrast, when sketch closely resembles multiple images, making its features challenging to distinguish with the naked eye, the model encounters errors that are justifiable. In addition, to test the generalization capability of OLTM on unknown styles, cross-style retrieval evaluation is performed on Market-Sketch-1K, and detailed results are in supplementary material.

Performance on PKU-Sketch. Tab. 2 presents the model's performance on PKU-SKetch dataset. The results indicate that our OLTM outperforms all competitors by a significant margin. For example, mAP and Rank-1 of OLTM are remarkably high at 91.4 and 94.0, surpassing state-of-the-art method by 5.2% and 4.0%, respectively. Because the sketches on PKU-Sketch contain more detailed information, they can assist multi-granularity interaction in acquiring more fine-grained knowledge. Fig. 3 illustrates the top-5 visualization results of OLTM on PKU-Sketch. Our method can accurately identify the target pedestrians despite challenges such as variations in posture, viewpoint, occlusion, sketch abstraction, and different painting styles.

5.3 Ablation Study

In this section, ablation experiments are conducted on Market-Sketch-1K to evaluate the effectiveness of each component within the OLTM framework.

The Effectiveness of Text Prompt Reasoning. To evaluate the contribution of different prompt setting, we train the model with different combination of each setting. As shown in Tab. 3, the combination of VQA and Prompt brings significant contribution. Compared to the strategy of directly aligning different modalities, leveraging TPR to implicitly guide the alignment results in improvements of 7.08% in mAP and 9.44% in rank-1. This significant improvement is primarily attributed to introducing textual information during the training phase, which enables the model to effectively capture semantic correlations between images and sketches during inference. Furthermore, compared to manually annotated fixed attributes, those generated by VQA enhance the model's performance by adaptively adjusting the level of detail on which it focuses. Moreover, the introduction of learnable prompts improved the mAP and rank-1 by 0.79% and 4.02%, respectively, compared to fixed templates. Prompt learning can enhance the network's learning and reasoning capabilities, allowing it to more flexibly adapt to diverse modalities and enhance its sensitivity to subtle distinctions.

The Effectiveness of Our Designed Modules. To validate the effectiveness of the TCA and CFI module, we progressively integrate them into the baseline and evaluate performance. The results in Tab. 3 indicate that both modules significantly enhance the alignment and interaction capabilities of model across modalities. Specifically, TCA improves mAP and rank-1 by 3.36% and 4.82% compared to baseline, respectively. The introduction of textual information in TCA effectively provides semantic guidance for coarse-grained alignment between modalities, enabling the model

to focus more on similar semantic relationships. Furthermore, the integration of CFI has increased mAP and rank-1 by 1.45% and 3.82%, respectively. CFI selectively focuses on key regions in visual concept representations through semantic consensus. This process optimizes feature interaction and ensures capturing more detailed information at a fine-grained level.

The Effectiveness of Triplet Assignment Loss. The experimental results in Tab. 3 demonstrate that the combination of proposed TAL $\mathcal{L}_{tal}$ and identity loss $\mathcal{L}_{ID}$ achieves optimal performance. The identity loss ensures that the model correctly identifies different individual identities, while TAL further optimizes feature space by pulling positive samples closer and pushing negative samples apart. Additionally, to verify the generalization of TAL, as shown in Tab. 4, we achieve superior performance across various network frameworks by substituting TAL for the weighted regularization triplet loss (WRT) [59] and hard triplet loss (HTL). Balancing between Euclidean distance and optimal transport distance can significantly enhance model performance. Please refer to the supplementary materials for more verification experiments on the key role of overall data distribution in enhancing sample feature distance.

Table 4: Performance of TAL $\mathcal{L}_{tal}$ with various baselines. '+' represents WRT; '*' represents HTL $\mathcal{L}_{htl}$.

Methods	mAP	R@1
BDG$^+$	24.45	24.70
BDG + TAL	**27.79**	**27.71**
baseline*	57.74	60.84
baseline + TAL	**58.41**	**61.04**
OLTM*	61.63	66.06
OLTM + TAL	**62.55**	**69.48**

6 Conclusion

In this paper, we present a optimal transport-based labor-free text prompt modeling (OLTM) framework for sketch re-identification. OLTM embeds text prompt reasoning module and distance measurement into transformer for achieving hierarchical multi-granularity alignment through the guidance of text semantics, leveraging the advantages of prompt learning and optimal transport. In addition, to address the limitations of Euclidean distance in measuring sample similarity, we propose a triplet assignment loss that guarantees a more effective standard based on the overall data distribution. Extensive experiments conducted on two public datasets indicate outstanding performance compared to other state-of-the-art methods for sketch Re-ID.

Discussion: In this work, we employ text injection and sample distance optimization to direct the model's attention toward key details, thereby minimizing performance losses due to modal gap and sample abstraction. However, our experiments revealed that when confronted with extremely vague sketch samples (i.e., those that human cannot discern features or match), the model's identification process deteriorates into a random selection among multiple potential outcomes. Therefore, sketch Re-ID heavily depend on the quality of sketches. Enhancing the discriminability of sketches without incurring additional labor costs is a topic worthy of further exploration in future research.

Acknowledgments

References

[1] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pages 1116–1124, 2015.

[2] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019.

[3] Jing Xu, Rui Zhao, Feng Zhu, Huaming Wang, and Wanli Ouyang. Attention-aware compositional network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2119–2128, 2018.

[4] Lu Pang, Yaowei Wang, Yi-Zhe Song, Tiejun Huang, and Yonghong Tian. Cross-domain adversarial feature learning for sketch re-identification. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 609–617, 2018.

[5] Cuiqun Chen, Mang Ye, Meibin Qi, and Bo Du. Sketchtrans: Disentangled prototype learning with transformer for sketch-photo recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[6] Kejun Lin, Zhixiang Wang, Zheng Wang, Yinqiang Zheng, and Shin'ichi Satoh. Beyond domain gap: Exploiting subjectivity in sketch-based person retrieval. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 2078–2089, 2023.

[7] Cuiqun Chen, Mang Ye, and Ding Jiang. Towards modality-agnostic person re-identification with descriptive query. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15128–15137, 2023.

[8] Shaojun Gui, Yu Zhu, Xiangxiang Qin, and Xiaofeng Ling. Learning multi-level domain invariant features for sketch re-identification. *Neurocomputing*, 403:294–303, 2020.

[9] Yafei Zhang, Yongzeng Wang, Huafeng Li, and Shuang Li. Cross-compatible embedding and semantic consistent feature construction for sketch re-identification. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3347–3355, 2022.

[10] Cuiqun Chen, Mang Ye, Meibin Qi, and Bo Du. Sketch transformer: Asymmetrical disentanglement learning from dynamic synthesis. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4012–4020, 2022.

[11] Xingyu Liu, Xu Cheng, Haoyu Chen, Hao Yu, and Guoying Zhao. Differentiable auxiliary learning for sketch re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 3747–3755, 2024.

[12] Yajing Zhai, Yawen Zeng, Da Cao, and Shaofei Lu. Trireid: Towards multi-modal person re-identification via descriptive fusion model. In *Proceedings of the 2022 International Conference on Multimedia Retrieval*, pages 63–71, 2022.

[13] Enhao Ning, Changshuo Wang, Huang Zhang, Xin Ning, and Prayag Tiwari. Occluded person re-identification with deep learning: a survey and perspectives. *Expert Systems with Applications*, page 122419, 2023.

[14] Leonid V Kantorovich. On the translocation of masses. *Journal of mathematical sciences*, 133(4):1381–1382, 2006.

[15] Wanxing Chang, Ye Shi, Hoang Tuan, and Jingya Wang. Unified optimal transport framework for universal domain adaptation. *Advances in Neural Information Processing Systems*, 35:29512–29524, 2022.

[16] Bharath Bhushan Damodaran, Benjamin Kellenberger, Rémi Flamary, Devis Tuia, and Nicolas Courty. Deepjdot: Deep joint distribution optimal transport for unsupervised domain adaptation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 447–463, 2018.

[17] Kilian Fatras, Thibault Séjourné, Rémi Flamary, and Nicolas Courty. Unbalanced minibatch optimal transport; applications to domain adaptation. In *International Conference on Machine Learning*, pages 3186–3197. PMLR, 2021.

[18] Wanxing Chang, Ye Shi, and Jingya Wang. Csot: Curriculum and structure-aware optimal transport for learning with noisy labels. *Advances in Neural Information Processing Systems*, 36:8528–8541, 2023.

[19] Chuanwen Feng, Yilong Ren, and Xike Xie. Ot-filter: An optimal transport filter for learning with noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16164–16174, 2023.

[20] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947, 2020.

[21] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8922–8931, 2021.

[22] Chao Zhang, Stephan Liwicki, and Roberto Cipolla. Beyond the cls token: Image reranking using pretrained vision transformers. In *BMVC*, page 80, 2022.

[23] Sergio Izquierdo and Javier Civera. Optimal transport aggregation for visual place recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16164–16174, 2023.

[24] Jiangming Shi, Yachao Zhang, Xiangbo Yin, Yuan Xie, Zhizhong Zhang, Jianping Fan, Zhongchao Shi, and Yanyun Qu. Dual pseudo-labels interactive self-training for semi-supervised visible-infrared person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11218–11228, 2023.

[25] Jiangming Wang, Zhizhong Zhang, Mingang Chen, Yi Zhang, Cong Wang, Bin Sheng, Yanyun Qu, and Yuan Xie. Optimal transport for label-efficient visible-infrared person re-identification. In *European Conference on Computer Vision*, pages 93–109. Springer, 2022.

[26] Yongguo Ling, Zhun Zhong, Zhiming Luo, Fengxiang Yang, Donglin Cao, Yaojin Lin, Shaozi Li, and Nicu Sebe. Cross-modality earth mover's distance for visible thermal person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 1631–1639, 2023.

[27] Zhizhong Zhang, Yuan Xie, Ding Li, Wensheng Zhang, and Qi Tian. Learning to align via wasserstein for person re-identification. *IEEE Transactions on Image Processing*, 29:7104–7116, 2020.

[28] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021.

[29] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021.

[30] Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Lam Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *arXiv preprint arXiv:2110.07602*, 2021.

[31] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16816–16825, 2022.

[32] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.

[33] Beier Zhu, Yulei Niu, Yucheng Han, Yue Wu, and Hanwang Zhang. Prompt-aligned gradient for prompt tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15659–15669, 2023.

[34] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727. Springer, 2022.

[35] Zifeng Wang, Zizhao Zhang, Sayna Ebrahimi, Ruoxi Sun, Han Zhang, Chen-Yu Lee, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, et al. Dualprompt: Complementary prompting for rehearsal-free continual learning. In *European Conference on Computer Vision*, pages 631–648. Springer, 2022.

[36] Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 139–149, 2022.

[37] Bingzhi Chen, Zhongqi Wu, Yishu Liu, Biqing Zeng, Guangming Lu, and Zheng Zhang. Enhancing cross-modal retrieval via visual-textual prompt hashing. In *Proceedings of international joint conference on artificial intelligence (IJCAI)*, pages 1–10, 2024.

[38] Rajshekhar Das, Yonatan Dukler, Avinash Ravichandran, and Ashwin Swaminathan. Learning expressive prompting with residuals for vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3366–3377, 2023.

[39] Aneeshan Sain, Ayan Kumar Bhunia, Pinaki Nath Chowdhury, Subhadeep Koley, Tao Xiang, and Yi-Zhe Song. Clip for all things zero-shot sketch-based image retrieval, fine-grained or not. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2765–2775, 2023.

[40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[41] Yajing Zhai, Yawen Zeng, Zhiyong Huang, Zheng Qin, Xin Jin, and Da Cao. Multi-prompts learning with cross-modal alignment for attribute-based person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 6979–6987, 2024.

[42] He Li, Mang Ye, Ming Zhang, and Bo Du. All in one framework for multimodal re-identification in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17459–17469, 2024.

[43] Daeha Kim and Byung Cheol Song. Optimal transport-based identity matching for identity-invariant facial expression recognition. *Advances in Neural Information Processing Systems*, 35:18749–18762, 2022.

[44] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

[45] Jason Altschuler, Jonathan Niles-Weed, and Philippe Rigollet. Near-linear time approximation algorithms for optimal transport via sinkhorn iteration. *Advances in neural information processing systems*, 30, 2017.

[46] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, Mingliang Xu, and Yi-Dong Shen. Dual-path convolutional image-text embeddings with instance loss. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 16(2):1–23, 2020.

[47] Peng Wu, Xuerong Zhou, Guansong Pang, Lingru Zhou, Qingsen Yan, Peng Wang, and Yanning Zhang. Vadclip: Adapting vision-language models for weakly supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 6074–6082, 2024.

[48] Haiwen Diao, Ying Zhang, Lin Ma, and Huchuan Lu. Similarity reasoning and filtration for image-text matching. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 1218–1226, 2021.

[49] Jianfeng Dong, Xirong Li, Chaoxi Xu, Xun Yang, Gang Yang, Xun Wang, and Meng Wang. Dual encoding for video retrieval by text. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):4065–4080, 2021.

[50] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[51] Mang Ye, Jianbing Shen, David J. Crandall, Ling Shao, and Jiebo Luo. Dynamic dual-attentive aggregation learning for visible-infrared person re-identification. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*, pages 229–247. Springer, 2020.

[52] Chaoyou Fu, Yibo Hu, Xiang Wu, Hailin Shi, Tao Mei, and Ran He. Cm-nas: Cross-modality neural architecture search for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11823–11832, 2021.

[53] Mang Ye, Weijian Ruan, Bo Du, and Mike Zheng Shou. Channel augmented joint learning for visible-infrared recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13567–13576, 2021.

[54] Yukang Zhang, Yan Yan, Yang Lu, and Hanzi Wang. Towards a unified middle modality learning for visible-infrared person re-identification. In *Proceedings of the 29th ACM international conference on multimedia*, pages 788–796, 2021.

[55] Mouxing Yang, Zhenyu Huang, Peng Hu, Taihao Li, Jiancheng Lv, and Xi Peng. Learning with twin noisy labels for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14308–14317, 2022.

[56] Hanzhe Sun, Jun Liu, Zhizhong Zhang, Chengjie Wang, Yanyun Qu, Yuan Xie, and Lizhuang Ma. Not all pixels are matched: Dense contrastive learning for cross-modality person re-identification. In *Proceedings of the 30th ACM international conference on multimedia*, pages 5333–5341, 2022.

[57] Yiyuan Zhang, Yuhao Kang, Sanyuan Zhao, and Jianbing Shen. Dual-semantic consistency learning for visible-infrared person re-identification. *IEEE Transactions on Information Forensics and Security*, 18:1554–1565, 2022.

[58] Yukang Zhang and Hanzi Wang. Diverse embedding expansion network and low-light cross-modality benchmark for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2153–2162, 2023.

[59] 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 transactions on pattern analysis and machine intelligence*, 44(6):2872–2893, 2021.

[60] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*, pages 5583–5594. PMLR, 2021.

[61] Chuhao Zhou, Jinxing Li, Huafeng Li, Guangming Lu, Yong Xu, and Min Zhang. Video-based visible-infrared person re-identification via style disturbance defense and dual interaction. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 46–55, 2023.

[62] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[63] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[64] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[65] Qian Yu, Feng Liu, Yi-Zhe Song, Tao Xiang, Timothy M Hospedales, and Chen-Change Loy. Sketch me that shoe. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 799–807, 2016.

[66] Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. The sketchy database: learning to retrieve badly drawn bunnies. *ACM Transactions on Graphics (TOG)*, 35(4):1–12, 2016.

[67] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

[68] Brandon Amos and J Zico Kolter. Optnet: Differentiable optimization as a layer in neural networks. In *International Conference on Machine Learning*, pages 136–145. PMLR, 2017.

[69] Yang Qin, Yingke Chen, Dezhong Peng, Xi Peng, Joey Tianyi Zhou, and Peng Hu. Noisy-correspondence learning for text-to-image person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1–18, 2024.

[70] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

[71] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*, 2022.

A Details about Optimal Transport

This section mainly introduces optimal transport and its corresponding algorithm, Sinkhorn-Knopp [44]. Let $\Gamma_r := \{x \in \mathbb{R}_+^r | x^\top 1_r = 1\}$ represents the probability simplex, where 1_r is the r-dimensional vector of ones. Given two probability simplex vectors $\alpha \in \Gamma_m$ and $\beta \in \Gamma_n$ and a cost matrix $C \in \mathbb{R}^{m \times n}$, the objective of OT is to seek the optimal transport plan P^* mapping α to β at the minimum cost:

$$
d_C(\alpha, \beta) = \min_{P \in U(\alpha,\beta)} \langle C, P \rangle,
$$
$$
U(\alpha, \beta) = \left\{ P \in \mathbb{R}_+^{m \times n} \mid P1_n = \alpha, P^\top 1_m = \beta \right\}, \tag{7}
$$

where $U(\alpha, \beta)$ denotes the transport polytope of α and β, i.e., the solution space of P. The above problem is to find optimal solution P^* in a set of all possible joint probabilities of (X, Y), where X and Y represent random variables with marginal distribution α and β.

Generally, optimal transport (OT) problem is a linear programming problem that can theoretically be solved in polynomial time. However, in the actual solving process, it involves the square of anchor feature dimensions at all scales, requiring near-cubic complexity [68]. Therefore, consider optimizing the problem through iteration. This method optimizes the solving process by adding an entropy constraint to the OT problem, transforming Eq. 7 into a non-linear convex form with a regularization term:

$$
d_C(\alpha, \beta) = \min_{P \in U(\alpha,\beta)} \langle C, P \rangle + \delta E(P),
$$
$$
U(\alpha, \beta) = \left\{ P \in \mathbb{R}_+^{m \times n} \mid P1_n = \alpha, P^\top 1_m = \beta \right\}, \tag{8}
$$

where δ is a hyper-parameter and $E(P) = P(\log P - 1)$ is the entropy of P. Introducing a regularization term is equivalent to introducing a prior knowledge: without considering the cost matrix C, the distribution of assignment matrix P is expected to be as uniform as possible. Eq. 8 is an entropy-regularized OT (EOT) problem and can be solved by Sinkhorn-Knopp algorithm [44] based on iterative updates of vectors. According to the Lagrange Multiplier Method, the conditional extremum problem (i.e., Eq. 8) can be transformed into an unconditional extremum problem:

$$
\hat{d}_C(\alpha, \beta) = \min_P \langle C, P \rangle + \delta E(P) + \mu(P1_n - \alpha) + \rho(P^\top 1_m - \beta) \tag{9}
$$

where, μ and ρ are the Lagrange multipliers. If we take its derivative and set it to 0, we can find P^*:

$$
P^* = exp(-\frac{\mu}{\delta})exp(-\frac{C}{\delta})exp(-\frac{\rho}{\delta}) \tag{10}
$$

Let $u = exp(-\frac{\mu}{\delta})$, $v = exp(-\frac{\rho}{\delta})$ and $K = exp(-\frac{C}{\delta})$, two constraint of Eq. 8 needs to be met simultaneously. Thus, one possible solution is to iterate enough times according to the following iteration formula:

$$
u^z = \frac{\alpha}{Kv^{z-1}}, v^z = \frac{\beta}{K^\top u^z}. \tag{11}
$$

Eq. 11 is called as Sinkhorn-Knopp iteration. After z iterations, P^* can be obtained by the following equation:

$$
P^* = Diag(u)KDiag(v) \tag{12}
$$

It is worth noting that the process of updating u and v (i.e., Eq. 11) alternately can be simplified to a single step: $u^z \leftarrow \alpha/K(\beta/K^\top u^{z-1})$. Furthermore, when all elements of α are positive, this single step can be further simplified as: $u^z \leftarrow 1./\hat{K}(\beta/K^\top u^{z-1})$, where $\hat{K} = Diag(1./\alpha)K$. The overall flow of Sinkhorn-Knopp is described by Algorithm 1.

B Details about Triplet Assignment Loss

B.1 Derivation for Gradient

This appendix provides some details on gradient derivation. To simplify representation and analysis, we focus on a single direction, following the approach in [69], considering that sketch-to-image retrieval and image-to-sketch retrieval are symmetrical. Moreover, we assume that there is only one

119213

Algorithm 1 The OT problem via Sinkhorn-Knopp

Require: Cost matrix C, weights α, β
Require: Hyper-parameter δ, iterations $i = 1$, max iteration z
1: Initialize $K = e^{-C/\delta}$, $\hat{K} = Diag(1./\alpha)K$, $u_0 = 1_n/n$
2: **while** u changes or i is less than or equal z **do**
3:    $u^i \leftarrow 1./\hat{K}(\beta/K^\top u^{i-1})$
4:    $i \leftarrow i + 1$
5: **end while**
6: Get optimal value $u^* = u$
7: $v^* = \beta/K^\top u^*$
8: **return** optimal flow matrix $P^* = Diag(u^*)KDiag(v^*)$

---

paired image for each sketch in the mini-batch. Consequently, we can simplify TRL and TAL as shown below:

$$\mathcal{L}_{trl}(R_i, S_i) = [m - r_i^\top s_i + \hat{r}_i^\top s_i]_+,$$

$$\mathcal{L}_{tal}(R_i, S_i) = [m - d(r_i, s_i) + d(\hat{r}_i, s_i))]_+, \; d(r,s) = \gamma r^\top s + (1-\gamma)(1 - exp(-\tfrac{r^\top s}{\delta}))r^\top s \tag{13}$$

where δ is the hyper-parameter of OT problem, $\hat{r}_i$ and r_i are the hardest negative sample and positive sample of the anchor sample s_i, respectively. These l_2-normalized features are embedded by the modality-specific models, i.e., $f_{\theta_r}(\cdot)$ and $f_{\theta_s}(\cdot)$. Due to the truncation operation $[x]_+$, we only discuss the case of $\mathcal{L} > 0$ that could generate gradients. For TRL, the gradients to the parameters θ_r and θ_s are:

$$\frac{\partial \mathcal{L}_{trl}}{\partial \theta_s} = \frac{\partial \mathcal{L}_{trl}}{\partial s_i}\frac{\partial s_i}{\partial \theta_s}, \quad \frac{\partial \mathcal{L}_{trl}}{\partial \theta_r} = \frac{\partial \mathcal{L}_{trl}}{\partial r_i}\frac{\partial r_i}{\partial \theta_r} + \frac{\partial \mathcal{L}_{trl}}{\partial \hat{r}_i}\frac{\partial \hat{r}_i}{\partial \theta_r}. \tag{14}$$

Since the learning of normalized features can be viewed as the movement process of points on a unit hyperplane, we only consider the loss gradients with respect to r_i, $\hat{r}_i$ and s_i are:

$$\frac{\partial \mathcal{L}_{trl}}{\partial r_i} = -s_i, \quad \frac{\partial \mathcal{L}_{trl}}{\partial s_i} = \hat{r}_i - r_i, \quad \frac{\partial \mathcal{L}_{trl}}{\partial \hat{r}_i} = s_i. \tag{15}$$

For our TAL, the gradients to the parameters θ_r and θ_s are:

$$\frac{\partial \mathcal{L}_{tal}}{\partial \theta_s} = \frac{\partial \mathcal{L}_{tal}}{\partial s_i}\frac{\partial s_i}{\partial \theta_s}, \quad \frac{\partial \mathcal{L}_{tal}}{\partial \theta_r} = \frac{\partial \mathcal{L}_{tal}}{\partial r_i}\frac{\partial r_i}{\partial \theta_r} + \frac{\partial \mathcal{L}_{tal}}{\partial \hat{r}_i}\frac{\partial \hat{r}_i}{\partial \theta_r}. \tag{16}$$

Since the result of $exp(-\tfrac{r^\top s}{\delta})$ is obtained through several iterations of Sinkhorn-Knopp algorithm, this part does not conduct gradients and can be simplified as a coefficient $\hat{\delta} = 1 - exp(-\tfrac{r^\top s}{\delta}) \in [0, 1]$. Thus, the gradients for r_i, $\hat{r}_i$ and s_i are:

$$\frac{\partial \mathcal{L}_{tal}}{\partial r_i} = -[\gamma + (1-\gamma)\hat{\delta}]s_i, \quad \frac{\partial \mathcal{L}_{tal}}{\partial \hat{r}_i} = [\gamma + (1-\gamma)\hat{\delta}]s_i,$$

$$\frac{\partial \mathcal{L}_{tal}}{\partial s_i} = \gamma(\hat{r}_i - r_i) + (1-\gamma)\hat{\delta}(\hat{r}_i - r_i). \tag{17}$$

B.2 The Effectiveness of Triplet Assignment Loss

We have included visual analysis in Fig. 4, which illustrates the convergence curve and sample distances during training. Figure 1(a) shows that conventional triplet loss converges prematurely. In contrast, our proposed triplet assignment loss exhibits higher volatility, reducing the risk of suboptimal local minima. Additionally, Figure 1(b) shows that a specific sketch sample (red box in the top left image) may have similar Euclidean distances to multiple RGB samples. The triplet assignment loss comprehensively considers the distribution of all samples (red box in the lower right image), offering broader possibilities for selecting the most relevant ones.

(a) The convergence curves of different losses　　　(b) The distance distribution across modalities
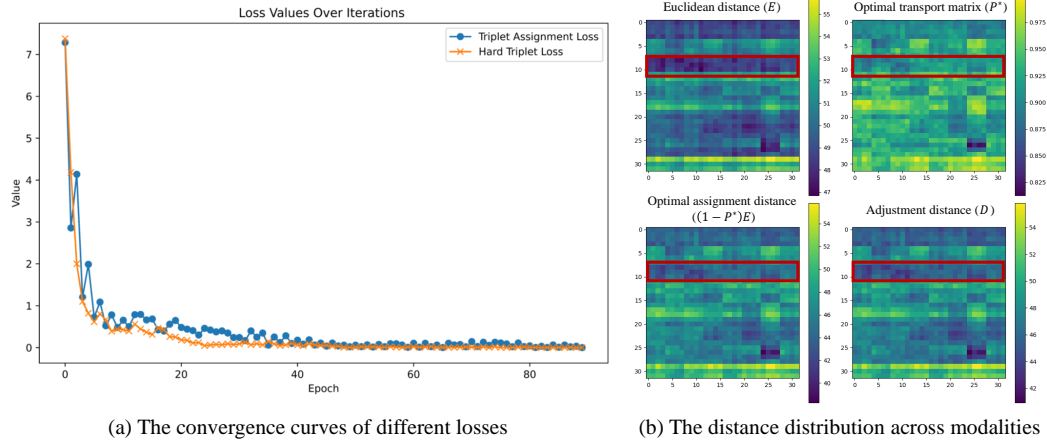
Figure 4: The effectiveness analysis of Triplet Assignment Loss. In Figure (b), the vertical axis represents RGB images, and the horizontal axis represents sketches. For each ID, 4 training examples are sampled, so the 4x4 cells on the diagonal represent positive sample pairs.

C　Computational Complexity

Our OLTM achieves the trade-off between performance enhancement and computational complexity. To this end, we select several methods for comparing parameters, floating-point operations (FLOPs), and frames per second (FPS), as shown in the Tab. 5 below. The results show that OLTM gets remarkable performance while maintaining reasonable computational costs. The reason is that: 1) only TCA module is required for inference; 2) the Visual Question Answering(VQA) model is used during data processing.

Table 5: The number of parameters, FLOPs, and FPS of different methods, where **bold** indicates the best performance in this field and <u>underline</u> indicates the second-best performance. VI denotes visible-infrared person re-identification.

Methods	Field	Backbone	Reference	Paras(M)	FLOPs(G)	FPS
DDAG[53]	VI	ResNet50	ECCV'2020	95.6	**5.2**	<u>13.2</u>
CM-NAS[54]	VI	ResNet50	ICCV'2021	**24.5**	**5.2**	14.8
CAJ[55]	VI	ResNet50	ICCV'2021	<u>71.6</u>	**5.2**	13.1
DEEN[56]	VI	ResNet50	CVPR'2023	89.3	<u>13.8</u>	7.7
CCSC[9]	Sketch	ViT	MM'2022	203.9	383.8	-
BDG[6]	Sketch	ResNet50	MM'2023	222.9	**5.2**	**14.9**
UNIReID[7]	Unified	CLIP	CVPR'2023	**149.6**	9.3	8.2
OLTM (Ours)	Sketch	CLIP	-	<u>181.9</u>	<u>6.2</u>	<u>11.7</u>

D　Additional Experimental Results

D.1　Experimental Setup

Implementation Details　The dimensions of image and text features are set to 512. Within a batch, we randomly select 8 identities, each comprising 4 images and 4 sketches. Each image is associated with 9 fine-grained textual attributes. To ensure more reliable comparisons, the random seeds are all set to 0. In the Text Prompt Reasoning module, ConMLP consists of a stack of $N = 2$ identical MLPs, where θ_{mlp} represents the trainable parameters. The iteration number of the Optimal Transport algorithm is 3. In the Triplet Assignment Loss, the iteration number of the Optimal Transport algorithm is 50. The model is trained with the Adam optimizer, starting with a learning rate of 1e-5, decaying with a cosine scheduler. The model is implemented in PyTorch on the RTX 4090 24GB GPU.

D.2 Experiment Result

Cross-style Retrieval Given the extensive Market-Sketch-1K dataset, where each person is sketched by six different artists, notable variations exist across these sketches. Consequently, we devise this experiment to assess our model's resilience to diverse artistic styles. Experimental setups involve sketches labeled as S1 to S6, each originating from different artists. Models are trained on sketches by specific artists and tested on sketches by others. And, "single query" denotes separate queries for sketches by different artists of the same individual, while "multi query" indicates queries combining multiple sketches of the same person. The specific details of these experiments are presented in Tab. 6.

Table 6: Testing on unseen styles. We report the mAP score.

a) single-query train and single-query test

train \ test	S_6	$S_{5,6}$	$S_{4...6}$	$S_{3...6}$	$S_{2...6}$
S_1	25.83	22.41	22.71	23.47	22.96
$S_{1,2}$	27.10	22.72	23.18	26.47	–
$S_{1...3}$	25.75	26.60	26.83	–	–
$S_{1...4}$	32.60	30.07	–	–	–
$S_{1...5}$	33.87	–	–	–	–

b) multi-query train and multi-query test

train \ test	S_6	$S_{5,6}$	$S_{4...6}$	$S_{3...6}$	$S_{2...6}$
S_1	–	–	–	–	–
$S_{1,2}$	–	31.82	33.68	39.49	–
$S_{1...3}$	–	34.30	36.49	–	–
$S_{1...4}$	–	39.42	–	–	–
$S_{1...5}$	–	–	–	–	–

c) single-query train and multi-query test

train \ test	S_6	$S_{5,6}$	$S_{4...6}$	$S_{3...6}$	$S_{2...6}$
S_1	–	29.17	30.35	34.65	34.49
$S_{1,2}$	–	31.23	34.34	41.05	–
$S_{1...3}$	–	34.52	37.65	–	–
$S_{1...4}$	–	40.10	–	–	–

d) multi-query train and single-query test

train \ test	S_6	$S_{5,6}$	$S_{4...6}$	$S_{3...6}$	$S_{2...6}$
$S_{1,2}$	27.10	22.68	23.35	24.81	–
$S_{1...3}$	26.97	25.13	31.96	–	–
$S_{1...4}$	29.99	27.79	–	–	–
$S_{1...5}$	31.96	–	–	–	–

Multi-query Setting Similar to [6], "multi-query" involves combining multiple sketches of the same ID during both training and inference. Our paper employs a straightforward fusion method by averaging the image features from multiple sketches. Tab. 7 below provides a comparative analysis of various fusion strategies. The results demonstrate that the basic and simple fusion method achieves the best experimental performance.

Table 7: Performance comparison of different multi-query experimental methods.

Methods	mAP	Rank@1	Rank@5	Rank@10
Simple Fusion	**62.55**	**69.48**	**90.36**	**95.18**
Average Pooling	60.95	66.27	88.15	94.38
Non-local Attention	60.98	65.66	90.16	94.98

Overfitting Analysis To mitigate overfitting, we apply various data augmentation techniques, including random cropping, rotation, and style augmentation. Furthermore, to validate the model's robustness and generalization, we conduct supplementary evaluation experiments on two large-scale datasets (SYSU-MM01 and RegDB) for visible-infrared person re-ID, as shown in Tab. 8. The results demonstrate that our OLTM achieves comparable performance in the visible-infrared domain.

Parameter Analysis For the Triplet Assignment Loss proposed in our work, we compute the sample distance using Eq. 18. Fig. 5 illustrates an analysis of the hyper-parameter γ. It can be observed that setting $\gamma = 0.3$ yields the best performance in Rank-1 and mAP.

$$D(R_i, S_i) = \gamma E(R_i, S_i) + (1 - \gamma)(1 - P_{i,i}^*)E(R_i, S_i) \tag{18}$$

where $E(R_i, S_i) = \|R_i' - S_i'\|_2$ denotes the Euclidean distance between feature representations.

In Dynamic Consensus Acquisition, the cost matrix $C \in \mathbb{R}_+^{n \times m}$ can be calculated for assignment, where the (i, j)-th element $C_{i,j}$ represents the cost of assigning a feature to an atom. Concretely, to introduce certain priors as guidance [23], the enhanced features are used to learn the cost matrix through two fully connected layers initialized randomly. m is a hyper-parameter that needs to be set. Fig. 6 analyzes the values of m, showing that the best performance in terms of Rank-1 and mAP is achieved when $m = 32$.

Table 8: Comparison results of our method on visible-infrared datasets, namely SYSU-MM01 and RegDB. Market-Sketch-1K is a sketch dataset used for reference. "VI" and "Sketch" represent their respective task domains.

| Methods | Domain | SYSU-MM01 | | | | RegDB | | | | Market-Sketch-1K | |
		All Search		Indoor Search		VIS to IR		IR to VIS		Sketch to VIS	
		R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP
DDAG[53]	VI	54.8	53.0	61.0	68.0	69.3	63.5	68.1	61.8	11.2	12.1
CM-NAS[54]	VI	60.8	58.9	68.0	52.4	82.8	79.3	81.7	77.6	0.7	0.8
CAJ[55]	VI	69.9	66.9	76.3	80.4	85.0	79.1	84.8	77.8	1.5	2.4
DART[57]	VI	68.7	66.3	72.5	78.2	83.6	75.7	82.0	73.8	6.6	7.8
DCLNet[58]	VI	70.8	65.3	73.5	76.8	81.2	74.3	78.0	70.6	12.2	13.5
OLTM(Ours)	Sketch	70.6	68.6	76.2	80.4	84.8	77.2	83.9	75.5	36.8	38.4
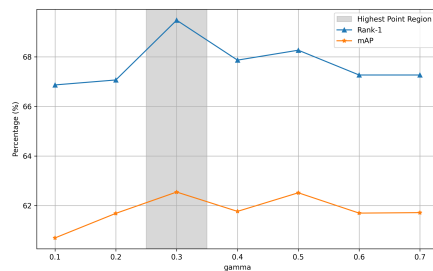
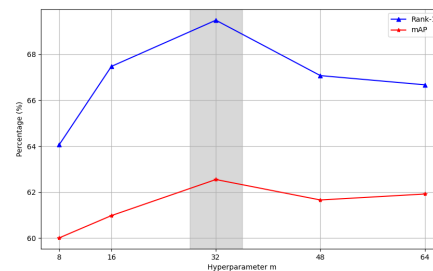Figure 5: Analysis of the hyperparameter γ

Figure 6: Analysis of the hyperparameter m

Qualitative Analysis Fig. 7 shows the Rank-5 results of OLTM and the baseline on the Market-Sketch-1K dataset. The left and right parts show the retrieval results of the baseline and OLTM, respectively. We can observe that OLTM can focus on more fine-grained discriminative information, such as bag and hat. In contrast, the baseline only considers global information matching, which leads to performance degradation.

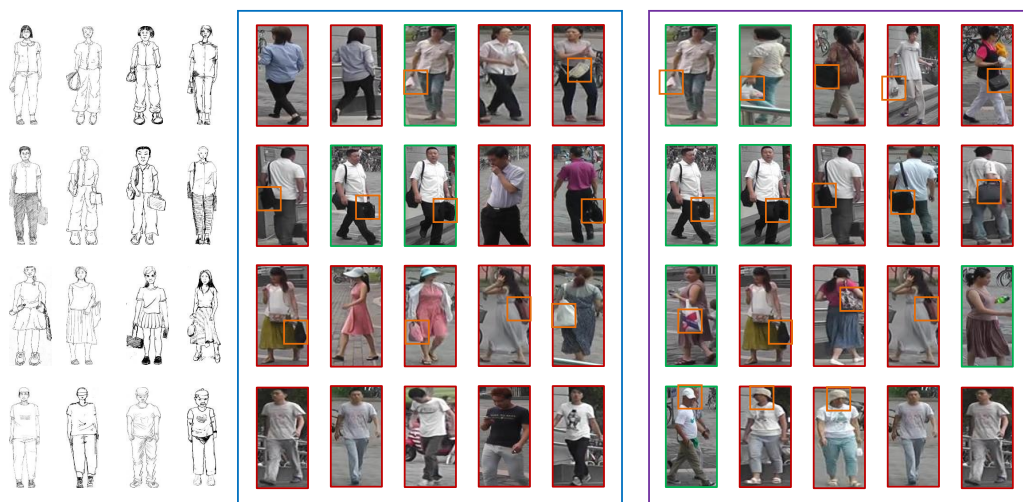Figure 7: The Rank-5 retrieval results under the multi-query setting on the Market-Sketch-1K dataset are presented. On the left side are the retrieval results of the baseline, and on the right side are the retrieval results of OLTM. Green borders indicate successful retrieval of the target pedestrian, while red borders indicate incorrect results. Yellow boxes represent fine-grained information.

E   Analysis of VQA Models

E.1   The Setup of the VQA Problems

Through extensive statistical analysis, we have formulated 9 specific questions to obtain the corresponding fine-grained textual attributes for a given image. These nine questions are as follows:

1. What is the gender of this person?
2. Is this person with long or short hair?
3. What is the color of this person's shirt?
4. Is this person wearing long sleeves or short sleeves?
5. Is this person wearing pants or a dress underneath?
6. What is the color of this person's lower garment?
7. Is this person carrying a backpack?
8. Is this person wearing a hat?
9. Is this person wearing glasses?

E.2   Replaceability of VQA Model

The VQA model is inherently substitutable. Essentially, any visual-language model which is capable of generating target attribute information from images can serve as an alternative. We also use other VQA models to demonstrate their substitutability, as shown in the Tab. 9 below.

Table 9: Performance comparison of different VQA models.

Methods	mAP	R@1	R@5	R@10
BLIP [70]	62.63	67.87	91.37	96.79
GIT [71]	62.33	68.47	90.76	96.59
VILT (ours)	62.55	69.48	90.36	95.18

E.3   The Fine-grained Recognition Ability of VQA Model

The VQA model has the capability to describe fine-grained recognition information for the following reasons: 1) The VQA model generates detailed attributes about various aspects of the pedestrian target (e.g., hair, backpack, hat), rather than relying on complete descriptive sentences. 2) We provide a visualization comparison, as shown in Fig. 8. This comparison demonstrates that using text attributes can guide effectively the attention of model.

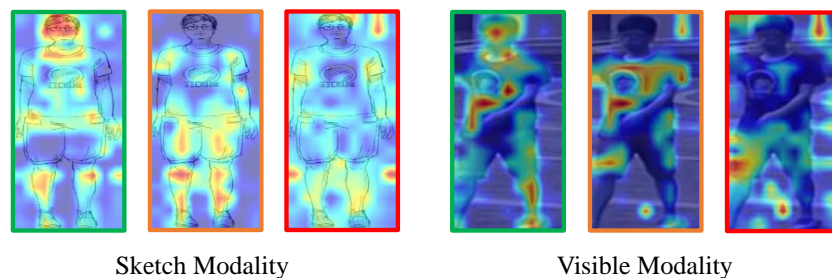Sketch Modality   Visible Modality

Figure 8: Visualization of attention maps. The **Green** and **Red** box indicate presence and absence of text attribute guidance, respectively. The **Orange** box represents partial text attribute guidance (excluding head-related information).

E.4 Background Information

We conduct additional background evaluation on Market-Sketch-1k dataset based on our initial studies. Specifically, we formulate the question *'What is the background of this image?'* to extract textual attributes about background. The extracted background details are illustrated in Fig. 9. However, the introduction of background information result in a decrease of 2.81 in Rank-1 and 1.62 in mAP. This decline can be attributed to the absence of corresponding background information in the sketches, which potentially interferes with the model's learning process.
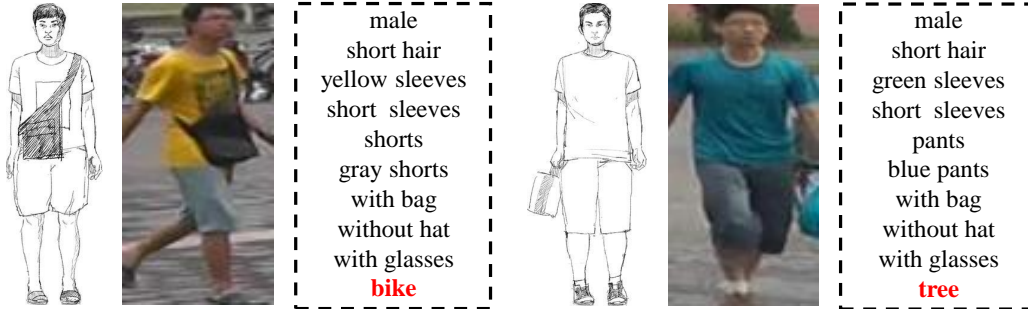

Figure 9: The text attributes generated by VQA model on RGB images. **Red** indicates background information obtained from the question: "What is the background of this image?".

E.5 The Availability of Text Attributes

To verify the effectiveness of different text attributes, we have provided additional ablation experiments in Tab. 10 below. The results show a significant decrease in model performance after discarding several hard-distinguished attributes (e.g., color and gender) in sketches. As shown in Fig. 10, sketches convey gender-related information through factors like body shape, and the contrast between light and dark areas effectively highlights specific color details. The TPR module injects detailed information into modal interactions during training. This enables the model to focus on these nuances autonomously, even without TPR during inference.

Table 10: Building on the original experimental setup, this comparison evaluates performance by removing the fine-grained textual attributes of gender, up color, and down color.

gender	up color	down color	mAP	Rank@1	Rank@5	Rank@10
✓	✓	✓	62.55	69.48	90.36	95.18
-	✓	✓	62.28	68.07	89.96	95.18
-	-	✓	61.64	67.67	89.76	94.98
-	-	-	61.35	67.07	89.16	94.78

Figure 10: Comparative analysis of RGB images and corresponding sketches. Gender-related factors in sketches include body shape and hair, and the contrast highlights color details.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction accurately reflect the paper's contributions and scope. The abstract and introduction provide a concise overview of the research area, the novel approaches introduced, and the key contributions, ensuring that readers have a clear understanding of what the paper aims to achieve and the significance of its findings.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: In the "Discussion" at the end of the paper, we discuss in detail the limitations of this work. *However, our experiments revealed that when confronted with extremely vague sketch samples (i.e., those that human cannot discern features or match), the model's identification process deteriorates into a random selection among multiple potential outcomes. Therefore, sketch Re-ID heavily depend on the quality of sketches. Enhancing the discriminability of sketches without incurring additional labor costs is a topic worthy of further exploration in future research.*

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.

- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: All formulas and proofs in the paper are numbered and cross-referenced. The paper provides complete proofs of the formulas, and for those appearing in the supplementary material, a brief proof sketch is provided in the main text. The proofs in the paper are rigorously reasoned, adhere to accepted mathematical principles, and do not omit any critical steps. The theorems and lemmas relied upon in the proofs are appropriately referenced and cross-referenced.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper fully discloses all the information needed to reproduce the main experimental results. The experimental section of the paper provides detailed descriptions of experimental setups, configurations, and other relevant details.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example

(a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.

(b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: This paper does not currently provide open-access code, but it is planned to be made public in the future.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The paper specifies all the training and test details necessary to understand the results. It includes information such as data splits, hyperparameters, how they were chosen, the type of optimizer used, and any other relevant details regarding the training and testing procedures.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.

- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The paper does not report error bars suitably and correctly defined, nor does it provide other appropriate information about the statistical significance of the experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: For each experiment, the paper provides sufficient information on the computer resources needed to reproduce the experiments (type of compute workers, memory).

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The paper discusses the potential positive societal impacts of the work. See introduction 1

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and the license and terms of use explicitly mentioned and properly respected.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: The paper does not release new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.