Unifying Multi-Modal Uncertainty Modeling and Semantic Alignment for Text-to-Image Person Re-identification

Zhiwei Zhao[1,2], **Bin Liu**[1,2*], **Yan Lu**[3], **Qi Chu**[1,2], **Nenghai Yu**[1,2]

[1]School of Cyber Science and Technology, University of Science and Technology of China
[2]CAS Key Laboratory of Electromagnetic Space Information
[3]Shanghai AI Laboratory
zwzhao98@mail.ustc.edu.cn, {flowice,qchu,ynh}@ustc.edu.cn, luyan@pjlab.org.cn

Abstract

Text-to-Image person re-identification (**TI-ReID**) aims to retrieve the images of target identity according to the given textual description. The existing methods in TI-ReID focus on aligning the visual and textual modalities through contrastive feature alignment or reconstructive masked language modeling (MLM). However, these methods parameterize the image/text instances as deterministic embeddings and do not explicitly consider the inherent uncertainty in pedestrian images and their textual descriptions, leading to limited image-text relationship expression and semantic alignment. To address the above problem, in this paper, we propose a novel method that unifies multi-modal uncertainty modeling and semantic alignment for TI-ReID. Specifically, we model the image and textual feature vectors of pedestrian as Gaussian distributions, where the multi-granularity uncertainty of the distribution is estimated by incorporating batch-level and identity-level feature variances for each modality. The multi-modal uncertainty modeling acts as a feature augmentation and provides richer image-text semantic relationship. Then we present a bi-directional cross-modal circle loss to more effectively align the probabilistic features between image and text in a self-paced manner. To further promote more comprehensive image-text semantic alignment, we design a task that complements the masked language modeling, focusing on the cross-modality semantic recovery of global masked token after cross-modal interaction. Extensive experiments conducted on three TI-ReID datasets highlight the effectiveness and superiority of our method over state-of-the-arts.

Introduction

Text-to-Image person re-identification (**TI-ReID**) is a subtask of person re-identification (Ye et al. 2022), aiming to retrieve the most matching pedestrian images from an image gallery based on given textual descriptions. Leveraging the ease of obtaining textual descriptions of the query compared to actual images, this technology offers a more versatile and user-friendly person search manner. Given its practical applicability in the domain of public safety, the TI-ReID has gained increasing attention in recent years.

Compared to general image-text retrieval, the TI-ReID is more challenging. It requires a fine-grained understanding

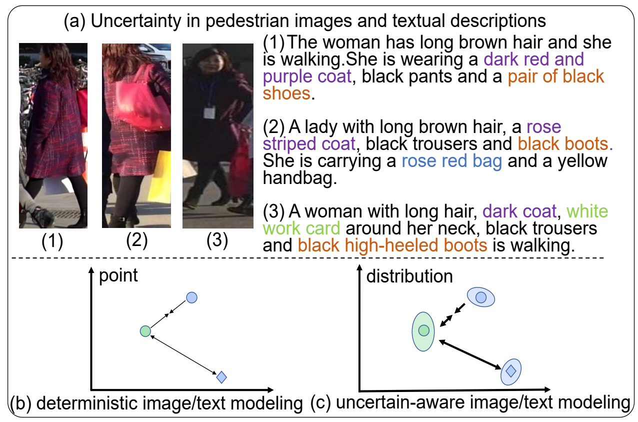

Figure 1: (a) The inherent uncertainty for pedestrian images and text descriptions in TI-ReID. (b) Current TI-ReID methods do not explicitly depict the uncertainty and parameterize visual-textual data as deterministic embeddings. (c) We model image/text embeddings as distributions and estimate the multi-granularity distribution uncertainty to express more reasonable and richer image-text relationships.

of the complex semantic concepts of pedestrians across the image and text modalities, as well as the establishment of cross-modal correspondences to bridge the inherent modality gap. Existing TI-ReID methods mainly revolve around aligning the image and text description of pedestrian into a shared space. They can be classified into cross-modal interaction-free (Zhang and Lu 2018; Han et al. 2021; Sarafianos, Xu, and Kakadiaris 2019; Wang et al. 2020) and cross-modal interaction-based (Li et al. 2017; Niu et al. 2020; Gao et al. 2021; Jiang and Ye 2023) methods. The former primarily utilized contrastive alignment (Zhang and Lu 2018; Han et al. 2021) to embed image-text features into shared space. In contrast, the latter employed the cross-attention mechanism (Niu et al. 2020; Farooq et al. 2022) and masked language modeling (Jiang and Ye 2023) to build fine-grained correlation between image regions and textual entities.

While successful to some extent, these methods have not explicitly considered the inherent uncertainty between pedestrian images and textual descriptions. As shown in Fig. 1 (a), uncertainty in pedestrian images arises from factors like viewpoint variations, lighting changes, while in tex-

*Corresponding author.

tual descriptions, it stems from word synonymy and granularity of annotations. Furthermore, the intra-modal uncertainty results in the same identity being associated with multiple perspectives of textual descriptions. Actually, this uncertainty reflects a reasonable range of semantic variation for image and text. Neglecting such uncertainty limits the semantic understanding and alignment capabilities for complex image-text relationships. This motivates us to explicitly model and utilize the uncertainty inherent in visual-textual data. In view of this, in this paper, we propose a novel approach that unifies multi-modal uncertainty modeling and semantic alignment for text-to-image person Re-ID.

Specifically, we first propose the multi-modal uncertainty modeling (**MUM**) for TI-ReID that characterizes the global features of pedestrian images and textual descriptions as Gaussian distributions. For each modality, the MUM estimates the multi-granularity uncertainty of distribution by combining batch-level and identity-level feature variance. The batch-level variance generally provides a coarse-grained reflection of modality-level uncertainty, while the identity-level variance captures the scope of fine-grained semantic variation. Random sampling from these probabilistic distributions acts as a multi-modal feature augmentation, which effectively enhances the diversity of image-text features and enriches more reasonable and meaningful image-text semantic relationships during training phase.

After utilizing the multi-modal uncertainty modeling to convey more comprehensive semantic relationships, it is essential to further strengthen the capability of cross-modal semantic alignment. We first develop a bi-directional cross-modal circle loss (**cm-Circle**) to more effectively align the probabilistic image and text features sampled from the distributions. Our cm-Circle loss is built upon the circle loss (Sun et al. 2020) in image retrieval and focuses on optimizing the similarity of cross-modal pairs from text-to-image and image-to-text with a self-paced manner. It can adaptively strengthen the alignment for under-optimized image-text pairs and well preserve the intra-modality structures. In addition, considering the current MLM-based methods (Jiang and Ye 2023) only focusing on utilizing visual context to recover the vocabulary semantics of masked *local* text tokens, we devise a task to recover the cross-modal semantic of masked *global* text token after the cross-modal interaction. This task (termed **cm-GSR**) employs cross-modal contrastive reconstruction as a supervisory signal, complementing the MLM and promoting comprehensive image-text semantic alignment and interaction. The multi-modal uncertainty modeling and semantic alignment objectives are integrated into a unified framework for end-to-end optimization.

Our main contributions can be summarized as follows:

- We present multi-modal uncertainty modeling for text-to-image person Re-ID, which uses Gaussian distributions to depict image/text features and estimates the multi-granularity uncertainty. It acts as feature augmentation and conveys richer image-text relationship.
- To enhance comprehensive image-text semantic alignment, we present a bi-directional cross-modal circle loss to align probabilistic image and text features more effectively, and propose to recover cross-modal semantic of masked global text token after cross-modal interaction.
- We unify the multi-modal uncertainty modeling and semantic alignment into a joint learning framework. Extensive experiments on three text-to-image person Re-ID datasets show the effectiveness and superiority of our approach against the state-of-the-arts.

Related Work

Text-to-Image Person Re-identification

Current TI-ReID methods can be roughly classified into cross-modal interaction-based and interaction-free methods. The interaction-based methods (Li et al. 2017; Niu et al. 2020; Wang et al. 2020; Farooq et al. 2022; Yan et al. 2022; Jiang and Ye 2023) utilize attention mechanisms to build fine-grained cross-modal correspondences between image regions and textual entities. Niu *et al.* (Niu et al. 2020) leveraged cross-attention to conduct relation-guided alignment between image regions and textual phrases, sentences. Gao *et al.* (Gao et al. 2021) proposed a contextual non-local attention mechanism to align full-scale image and textual features. Jiang *et al.* (Jiang and Ye 2023) further designed the cross-modal interaction transformer and used the masked language modeling (MLM) task to achieve implicit fine-grained alignment. The cross-modal interaction-free methods (Zheng et al. 2020; Zhang and Lu 2018; Han et al. 2021; Wang et al. 2020) focus on upgrading model structures and designing contrastive-style loss functions to extract and align image-text representations. Benefiting from the advancements of vision-language pretraining (VLP) (Radford et al. 2021), the encoders for image and text modalities in TI-ReID have undergone upgrades, transitioning from ResNet (He et al. 2016) and BERT (Devlin et al. 2018) to CLIP-based encoders (Radford et al. 2021). The representative loss functions in TI-ReID include cross-modal projection matching (CMPM) loss (Zhang and Lu 2018), cross-modality contrastive loss (Han et al. 2021), similarity distribution matching (SDM) loss (Jiang and Ye 2023). Nevertheless, the above approaches fail to consider the inherent uncertainty in pedestrian images and their corresponding textual descriptions, leading to limited image-text understanding and alignment capability. Furthermore, the MLM-based method (Jiang and Ye 2023) solely focus on semantic recovery for masked local text tokens, disregarding the global masked token. Additionally, the contrastive-style losses overlook the varying learning difficulty among different cross-modal samples. In this work, we explicitly model the mutli-modal uncertainty and promote more effective semantic alignment for TI-ReID.

Uncertainty Modeling in Computer Vision

Uncertainty modeling, which aims to capture the intrinsic "randomness" in the data, has been receiving increasing attention in computer vision. In face recognition and person Re-ID, the DUL (Chang et al. 2020) and DistributionNet (Yu et al. 2019) employed Gaussian distributions to model face/person embeddings and used a learnable sub-network to estimate uncertainty, reflecting the quality of facial/person features. In domain generalization, the DSU (Li et al. 2022)

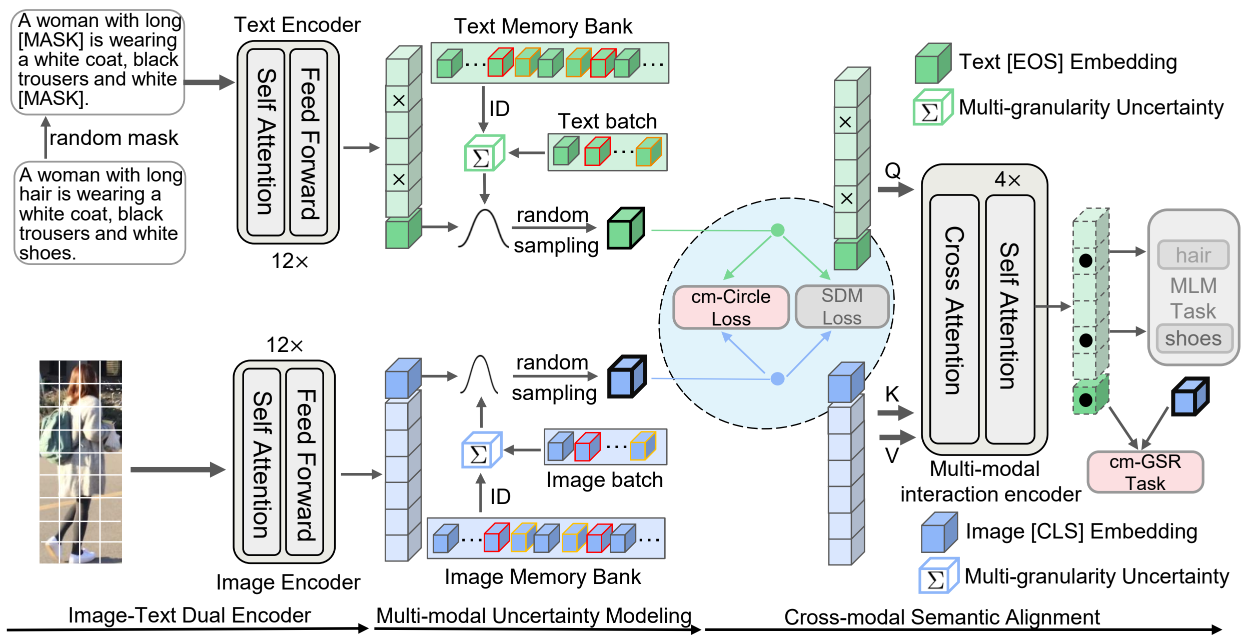

Figure 2: The overview framework of our proposed method for TI-ReID. Given the image and text inputs, we first present multi-modal uncertainty modeling to represent them as Gaussian distributions and estimate multi-granularity distribution uncertainty by jointly utilizing batch-level and identity-level feature variances. Subsequently, for further enhancing cross-modal semantic alignment, we propose the cross-modal circle loss (cm-Circle) to more effectively align the probabilistic image-text features in self-paced manner and present the cm-GSR task to promote more comprehensive image-text interaction and alignment.

modeled the uncertainty of feature statistics to generate diverse domain shifts. In cross-modality retrieval, the PCME (Chun et al. 2021) presented the probabilistic cross-modal embedding and predicted the mean and variance with learnable sub-networks. In vision-language pretraining, the MAP method (Ji et al. 2023) modeled image-text features as probabilistic distributions and utilizes learnable multi-head self-attention module to estimate uncertainty. In this paper, we present multi-modal uncertainty modeling for the first time in the text-to-image person Re-ID. We estimate the distribution uncertainty for each image/text instance with multi-granularities by jointly using batch-level and identity-level feature variances, which is more suitable for TI-ReID and expresses richer image-text semantic relationships.

Method

In this section, we present the joint multi-modal uncertainty modeling and semantic alignment method. An overview of the framework is illustrated in Figure 2 and we delve into its specific details in the following subsections.

Image-Text Dual Encoder

The inputs consist of image-text pairs, represented as $\{v_i, t_i, y_i\}_{i=1}^{B}$. where v_i, t_i, and y_i refer to the image, text, and identity label, respectively. B is the batch-size.

Image Encoder. We use a CLIP pre-trained Vision Transformer (ViT) to obtain the image embedding from an input image $v_i \in \mathbb{R}^{H\times W\times C}$. The image is split into a sequence of $N = H \times W/P^2$ patches, with P denoting the patch size. A trainable linear projection is applied to map these patches to 1D tokens $\{f_n^v\}_{n=1}^{N}$. The positional embeddings and [CLS] token are added to the token sequence. The resulting sequence of tokens is then processed through multiple layer transformer blocks to model relations between each patch and obtain the sequence of contextual image embeddings $\{f_{cls}^v, f_1^v, \cdots, f_N^v\}$. where the f_{cls}^v is served as the global image representation $\boldsymbol{g_i^v} \in \mathbb{R}^{512}$.

Text Encoder. For input text t_i, the CLIP text encoder is used to extract text representation. The text description is tokenized and enclosed with [SOS] and [EOS] tokens to indicate the sequence's beginning and end. Following recent methods (Shu et al. 2023; Wei et al. 2023), we randomly mask the word tokens of the input text t_i with a probability (usually 15% or 30%) and replace them with the special token [MASK] during training. The masked text sequence then fed into the transformer to obtain sequence of contextual text embedding $\{f_{sos}^t, f_1^t, \cdots, f_{eos}^t\}$, where the transformer uses masked self-attention to capture correlations among tokens. Finally, the embedding at the [EOS] token, f_{eos}^t is treated as the global text feature $\boldsymbol{g_i^t} \in \mathbb{R}^{512}$.

Multi-Modal Uncertainty Modeling

The inherent uncertainty of pedestrian images and textual descriptions reflects a reasonable range of semantic variation. This motivates us to explicitly model and utilize the uncertainty in visual-textual data. By employing this uncertainty for feature augmentation of visual-textual instance, it can effectively express more reasonable image-text semantic relationships and contribute diverse semantic alignment. We suggest that by incorporating potential uncertainties, the global features of each pedestrian image and textual description, conform to specific Gaussian distributions. Therefore, the key lies in efficiently and comprehensively estimating the uncertainty of distributions for pedestrian images and texts. We propose multi-modal uncertainty modeling (**MUM**), which estimates the uncertainty of distribution for image and text modalities by considering both the batch-level and identity-level feature variance. We believe that for each modality, the variance of feature embeddings within mini-batch primarily provides a coarse-grained perspective of image/text uncertainty and can be calculated by Eq. (1),

$$\begin{aligned}\Sigma^2_{\text{batch}}(\mathcal{V}) &= \frac{1}{B}\sum_{i=1}^{B}(\boldsymbol{g_i^v} - \mathbb{E}_b[\boldsymbol{g^v}])^2,\\ \Sigma^2_{\text{batch}}(\mathcal{T}) &= \frac{1}{B}\sum_{i=1}^{B}(\boldsymbol{g_i^t} - \mathbb{E}_b[\boldsymbol{g^t}])^2,\end{aligned} \tag{1}$$

where $\Sigma_{\text{batch}}(\mathcal{V}), \Sigma_{\text{batch}}(\mathcal{T}) \in \mathbb{R}^{512}$ represent the coarse-grained uncertainty for image/text modalities, respectively.

However, solely estimating modality-level coarse-grained uncertainty is insufficient for fine-grained TI-ReID task, we proceed to depict the important fine-grained uncertainty by considering the identity label. For visual and textual modalities, the identity-level feature variances are calculated to capture the local scope of semantic variations specific to the individuals. Given the difficulty of estimating identity-level variances through randomly sampled mini-batch, we employ two memory banks $\mathcal{M}^V$ and $\mathcal{M}^T$, composed of the first-in-first-out dynamic queues, to respectively record a significant amount of global visual and text features from image-text pairs in past and current iterations. Specifically, the $\mathcal{M}^V = \{\boldsymbol{h_i^v}\}_{i=1}^{|\mathcal{M}|}$ and the $\mathcal{M}^T = \{\boldsymbol{h_i^t}\}_{i=1}^{|\mathcal{M}|}$. The $\boldsymbol{h_i^v}$ and $\boldsymbol{h_i^t}$ are the global visual and textual features recorded in the memorys. The $|\mathcal{M}|$ denotes the size of memory bank, and is set to 65536. Then the identity-level feature variances for each identity in image and text modalities can be derived by

$$\begin{aligned}\Sigma^2_{\text{ID}}(\mathcal{V}_y) &= \frac{1}{|\mathcal{M}_y^V|}\sum_{i=1}^{|\mathcal{M}|}\mathbb{1}(y_i = y) * (\boldsymbol{h_i^v} - \mathbb{E}_m[\boldsymbol{h_y^v}])^2,\\ \Sigma^2_{\text{ID}}(\mathcal{T}_y) &= \frac{1}{|\mathcal{M}_y^T|}\sum_{i=1}^{|\mathcal{M}|}\mathbb{1}(y_i = y) * (\boldsymbol{h_i^t} - \mathbb{E}_m[\boldsymbol{h_y^t}])^2,\end{aligned} \tag{2}$$

where $\Sigma_{\text{ID}}(\mathcal{V}_y), \Sigma_{\text{ID}}(\mathcal{T}_y) \in \mathbb{R}^{512}$ indicate the fine-grained uncertainty of the y-th identity in vision and text modality, respectively. $\mathbb{1}(y_i = y)$ is the indicator function, $|\mathcal{M}_y^V|$ is the sample number of y-th identity in memory. We then unify the coarse-grained and fine-grained uncertainty through weighted coupling to estimate multi-granularity uncertainty $\Sigma_{\text{unify}}(\mathcal{V}_y)$ and $\Sigma_{\text{unify}}(\mathcal{T}_y)$ for each image/text instance by the Eq. (3), where the $\omega \in (0, 1)$ is the coupling factor and the s is the scale factor.

$$\begin{aligned}\Sigma_{\text{unify}}(\mathcal{V}_y) &= s * (\omega * \Sigma_{\text{batch}}(\mathcal{V}) + (1-\omega) * \Sigma_{\text{ID}}(\mathcal{V}_y)),\\ \Sigma_{\text{unify}}(\mathcal{T}_y) &= s * (\omega * \Sigma_{\text{batch}}(\mathcal{T}) + (1-\omega) * \Sigma_{\text{ID}}(\mathcal{T}_y)).\end{aligned} \tag{3}$$

The multi-granularity uncertainty $\Sigma_{\text{unify}}(\mathcal{V}_y)/\Sigma_{\text{unify}}(\mathcal{T}_y)$ not only captures modality-related coarse-grained global uncertainty patterns, but also encompasses fine-grained identity-related variations. With such multi-modal uncertainty modeling, it expands the reasonable and meaningful semantic distribution range for each visual/textual feature. Each visual/textual feature is established as Gaussian distribution with the uncertainty, and denoted as $\boldsymbol{p_i^v} \sim \mathcal{N}(\boldsymbol{g_i^v}, \Sigma^2_{\text{unify}}(\mathcal{V}_{y_i}))$ and $\boldsymbol{p_i^t} \sim \mathcal{N}(\boldsymbol{g_i^t}, \Sigma^2_{\text{unify}}(\mathcal{T}_{y_i}))$, respectively. Then the probabilistic features can be randomly drawn from the above distributions with the re-parameterization trick by as follows:

$$\begin{aligned}\boldsymbol{p_i^v} &= \boldsymbol{g_i^v} + \epsilon_i^v * \Sigma_{\text{unify}}(\mathcal{V}_{y_i}), \quad \epsilon_i^v \sim \mathcal{N}(\mathbf{0}, \mathbf{I}),\\ \boldsymbol{p_i^t} &= \boldsymbol{g_i^t} + \epsilon_i^t * \Sigma_{\text{unify}}(\mathcal{T}_{y_i}), \quad \epsilon_i^t \sim \mathcal{N}(\mathbf{0}, \mathbf{I}),\end{aligned} \tag{4}$$

where the ϵ_i^v and ϵ_i^t are individually sampled from standard normal distributions. By randomly sampling from the above distributions for each image/text instance, it can generate more reasonable features with different directions and intensities and express richer image-text semantic relationship.

Cross-Modal Semantic Alignment

After conveying richer visual-textual semantic relationships through the proposed multi-modal uncertainty modeling, we need to enhance the visual-textual semantic alignment to adapt more diverse features. We first employ the commonly used similarity distribution matching (SDM) loss (Jiang and Ye 2023) in TI-ReID to initially align the probabilistic image-text features. It minimizes the KL divergence between the distributions of image-text similarity $\pi_{i,j}$ and the normalized distributions of matching labels $q_{i,j}$ as follows:

$$\mathcal{L}^{t2v}_{\text{SDM}} = \frac{1}{B}\sum_{i=1}^{B}\sum_{j=1}^{B}\left(\pi_{i,j}\cdot\log\frac{\pi_{i,j}}{q_{i,j}+\delta}\right), \tag{5}$$

$$\pi_{i,j} = \frac{\exp\left(\boldsymbol{p_i^t}\cdot\boldsymbol{p_j^v}/\tau\right)}{\sum_{k=1}^{B}\exp\left(\boldsymbol{p_i^t}\cdot\boldsymbol{p_k^v}/\tau\right)}, q_{i,j} = \frac{l_{i,j}}{\sum_{k=1}^{B} l_{i,k}}, \tag{6}$$

where τ is temperature coefficient, the $\boldsymbol{p_i^t}\cdot\boldsymbol{p_v^j}$ is the cosine similarity. $l_{i,j} = 1$ means that (t_i, v_j) is positive pair with same identity, while $l_{i,j} = 0$ indicates negative pair, the δ is a small number to avoid the numerical issues. The total SDM loss $\mathcal{L}_{\text{SDM}} = \mathcal{L}^{t2v}_{\text{SDM}} + \mathcal{L}^{v2t}_{\text{SDM}}$.

To further enhance the semantic alignment between probabilistic global images and text features more efficiently, we present a bi-directional cross-modal circle loss (termed **cm-Circle**) for the TI-ReID, inspired by the circle loss (Sun et al. 2020) in image retrieval task. The designed cm-circle loss aims to further align the global semantic of probabilistic features for positive and negative image-text pairs in a self-paced manner. Specifically, the text-to-image cm-circle loss $\mathcal{L}^{t2v}_{\text{cmcir}}$ is formulated as Eq. (7), where the $\boldsymbol{p_i^t p_k^v}$ and $\boldsymbol{p_i^t p_j^v}$ denote the cosine similarity of positive and negative image-text pair in probabilistic feature space, respectively. The α_p^k and α_n^j respectively represent the non-negative re-weighting for each positive and negative image-text pair.

$$\mathcal{L}^{t2v}_{\text{cmcir}} = \log\left[1 + \sum_{j}^{y_i \neq y_j} e^{\gamma\alpha_n^j\left(\boldsymbol{p_i^t p_j^v} - \Delta_n\right)}\sum_{k}^{y_i = y_k} e^{-\gamma\alpha_p^k\left(\boldsymbol{p_i^t p_k^v} - \Delta_p\right)}\right] \tag{7}$$

Similarly, the image-to-text cm-circle loss $\mathcal{L}_{\text{cmcir}}^{v2t}$ can be expressed as Eq. (8) with a symmetric manner.

$$\mathcal{L}_{\text{cmcir}}^{v2t}=\log\left[1+\sum_{j}^{y_i\neq y_j} e^{\gamma\beta_n^j\left(\boldsymbol{p_i^v p_j^t}-\Delta_n\right)}\sum_{k}^{y_i=y_k} e^{-\gamma\beta_p^k\left(\boldsymbol{p_i^v p_k^t}-\Delta_p\right)}\right] \tag{8}$$

The re-weighting factors α_p^k, α_n^j and β_p^k, β_n^j are calculated as Eq. (9), where O_p and O_n are optimums of the similarity score for the positive and negative image-text pair, respectively. The hyper-parameter $O_p = 1 + m$, $O_n = -m$, $\Delta_p = 1 - m$, $\Delta_n = m$, m is the margin, $[\cdot]_+$=max$\{\cdot, 0\}$.

$$\begin{cases} \alpha_p^k = \left[O_p - \boldsymbol{p_i^t p_k^v}\right]_+, \;\; \beta_p^k = \left[O_p - \boldsymbol{p_i^v p_k^t}\right]_+, \\ \alpha_n^j = \left[\boldsymbol{p_i^t p_j^v} - O_n\right]_+, \;\; \beta_n^j = \left[\boldsymbol{p_i^v p_j^t} - O_n\right]_+. \end{cases} \tag{9}$$

Finally, the bi-directional cm-circle loss is calculated as Eq. (10), and it brings two benefits to TI-ReID. First, it focuses solely on optimizing similarity of cross-modal positive and negative text-image pairs, preserving intra-modality structures. Secondly, it dynamically adjusts the weights of cross-modal pairs based on alignment difficulty, enhancing optimization for under-optimized image-text pairs.

$$\mathcal{L}_{\text{cm-Circle}} = \mathcal{L}_{\text{cmcir}}^{t2v} + \mathcal{L}_{\text{cmcir}}^{v2t}. \tag{10}$$

The above proposed cross-modal circle loss only offers coarse-grained semantic alignment between vision and text. Following recent MLM-based method IRRA (Jiang and Ye 2023), as shown in Fig. 2, we use a multi-modal interaction encoder (MIE) consisting of several cross-attention and self-attention layers to model the interactions between the sequence of contextual image embeddings $f(v_i) = \{f_{cls}^v, f_1^v, \cdots, f_N^v\}$ and the sequence of contextual text embeddings $f(t_i) = \{f_{sos}^t, f_1^t, \cdots, f_{eos}^t\}$ of masked text. The $\{r_{i,k}^t\}_{k=1}^L$ denote the recovered textual token embeddings after cross-modal interaction, L is the length of text tokens.

$$\left\{r_{i,k}^t\right\}_{k=1}^L = \text{MIE}(f(t_i), f(v_i)). \tag{11}$$

Then, the masked language modeling predicts the correct vocabulary ID for masked word tokens with contextual image embeddings and textual embeddings, by minimizing the negative log-likelihood as Eq. (12). $\mathbb{M}_{\text{indexes}}$ is the indexes of masked positions, $w_{i,k}$ is the true vocabulary ID of word.

$$\mathcal{L}_{\text{MLM}} = -\mathbb{E}_{i,k\in\mathbb{M}_{\text{indexes}}} \log\, p(w_{i,k} \mid r_{i,k}^t). \tag{12}$$

However, we observe that above-mentioned masked language modeling task solely focus on recovering the vocabulary semantic for masked *local* text tokens, while ignoring the key *global* masked text embedding. Actually, the $\boldsymbol{r_{i,eos}^t}$ represents the recovered *global* embedding of masked text after the cross-modality interaction (via Eq. (11)). We encourage the $\boldsymbol{r_{i,eos}^t}$ should encompass complete cross-modal semantic, as achieving this objective necessitates a more comprehensive cross-modal interaction between $f(v_i)$ and $f(t_i)$. In view of this, we design a task (termed **cm-GSR**) to recover the cross-modal semantic of masked global text token after the cross-modal interaction, which leveraging the cross-modal contrastive reconstruction as supervisory signal. We apply the cross-modal Info-NCE loss between the $\boldsymbol{r_{i,eos}^t}$ and the complete image embedding $\boldsymbol{g_i^v}$ to achieve the cm-GSR task, and can be expressed as Eq. (15),

$$\mathcal{L}_{\text{NCE}}^{t2v} = -\mathbb{E}_i[\log \frac{\exp(<\boldsymbol{r_{i,eos}^t}, \boldsymbol{g_i^v}>/\tau)}{\sum_{j=1}^B \exp<\boldsymbol{r_{i,eos}^t}, \boldsymbol{g_j^v}>/\tau)}], \tag{13}$$

$$\mathcal{L}_{\text{NCE}}^{v2t} = -\mathbb{E}_i[\log \frac{\exp(<(\boldsymbol{g_i^v}, \boldsymbol{r_{i,eos}^t}>)/\tau)}{\sum_{j=1}^B \exp(<\boldsymbol{g_i^v}, \boldsymbol{r_{j,eos}^t}>)/\tau)}], \tag{14}$$

$$\mathcal{L}_{\text{cm-GSR}}(\boldsymbol{r_{i,eos}^t}, \boldsymbol{g_i^v}) = 0.5(\mathcal{L}_{\text{NCE}}^{t2v} + \mathcal{L}_{\text{NCE}}^{v2t}), \tag{15}$$

where $<\boldsymbol{r_{i,eos}^t}, \boldsymbol{g_i^v}>$ denotes the cosine similarity between $\boldsymbol{r_{i,eos}^t}$ and $\boldsymbol{g_i^v}$. The cm-GSR task effectively complements the masked language modeling and promotes more comprehensive image-text interaction and semantic alignment.

Joint Optimization

We unify multi-modal uncertainty modeling and cross-modal semantic alignment into an end-to-end framework, and minimize overall optimization loss $\mathcal{L}_{\text{overall}}$ for training.

$$\mathcal{L}_{\text{overall}} = \mathcal{L}_{\text{SDM}} + \mathcal{L}_{\text{MLM}} + \lambda_1\mathcal{L}_{\text{cm-Circle}} + \lambda_2\mathcal{L}_{\text{cm-GSR}}. \tag{16}$$

Experiments

Experimental Setup

CUHK-PEDES (Li et al. 2017) has 40,206 images and 80,412 textual descriptions associated with 13,003 identities. The training set has 11,003 identities with 34,054 images and 68,108 textual descriptions. The validation and test set comprise 3,078 and 3,074 images, along with 6,158 and 6,156 textual descriptions, respectively. Both the val/test subsets have 1,000 identities.

RSTPReid (Zhu et al. 2021) comprises 20,505 images, showcasing 4,101 unique identities. Each identity is represented by five images from different cameras, with every image being paired with two textual descriptions. The dataset utilizes 3,701, 200 and 200 identities for training, validation, and testing, respectively.

ICFG-PEDES (Ding et al. 2021) is a identity-centric TI-ReID dataset, featuring 54,522 images across 4,102 unique identities. Each image corresponds to a single textual description. The dataset is divided into a training set with 34,674 images from 3,102 identities and a test set containing 19,848 images representing 1,000 identities.

Evaluation Protocol. Similar to most works in TI-ReID, we report the Rank-k accuracy (k=1,5,10) and the mean Average Precision (mAP) metric.

Implementation Details. Our approach is implemented using the PyTorch framework on a single NVIDIA RTX-3090 GPU(24G). Similar to the IRRA method (Jiang and Ye 2023), our model comprises a pre-trained image encoder (CLIP-ViT-B/16), a pre-trained text encoder (CLIP text Transformer), and a randomly initialized multimodal interaction encoder. During training, all input images are resized to 384 × 128, the patch and stride size are set to 16. We apply the random horizontal flipping, RandAugment (Cubuk

Method	Venue	Image Encoder	Text Encoder	R1	R5	R10	mAP
GNA-RNN (Li et al. 2017)	CVPR17	VGG	LSTM	19.05	-	53.64	-
Dual-path (Zheng et al. 2020)	TOMM20	RN50	RN50	44.40	66.26	75.07	-
CMPM/C (Zhang and Lu 2018)	ECCV18	RN50	LSTM	49.37	-	79.27	-
MIA (Niu et al. 2020)	TIP20	RN50	GRU	53.10	75.00	82.90	-
PMA (Jing et al. 2020)	AAAI2020	RN50	LSTM	53.81	73.54	81.23	-
ViTAA (Wang et al. 2020)	ECCV20	RN50	LSTM	54.92	75.18	82.90	51.60
NAFS (Gao et al. 2021)	arXiv21	RN50	BERT	59.36	79.13	86.00	54.07
DSSL (Zhu et al. 2021)	MM21	RN50	BERT	59.98	80.41	87.56	-
SSAN (Ding et al. 2021)	arXiv21	RN50	LSTM	61.37	80.15	86.73	-
LapsCore (Wu et al. 2021)	ICCV21	RN50	BERT	63.40	-	87.80	-
TextReID (Han et al. 2021)	BMVC21	CLIP-RN101	CLIP-Xformer	64.08	81.73	88.19	60.08
TIPCB (Chen et al. 2022)	Neuro22	RN50	BERT	64.26	83.19	89.10	
CAIBC (Wang et al. 2022a)	MM22	RN50	BERT	64.43	82.87	88.37	-
AXM-Net (Farooq et al. 2022)	AAAI22	RN50	BERT	64.44	80.52	86.77	58.73
LGUR (Shao et al. 2022)	MM22	DeiT-Small	BERT	65.25	83.12	89.00	-
IVT (Shu et al. 2023)	ECCVW22	ViT-Base	BERT	65.59	83.11	89.21	-
CFine (Yan et al. 2022)	arXiv22	CLIP-ViT	BERT	69.57	85.93	91.15	-
MCM (Wei et al. 2023)	arXiv23	CLIP-ViT	CLIP-Xformer	69.61	86.01	90.90	-
IRRA (Jiang and Ye 2023)	CVPR2023	CLIP-ViT	CLIP-Xformer	73.38	89.93	93.71	66.13
baseline (CLIP-ViT-B/16)	-	CLIP-ViT	CLIP-Xformer	72.98	89.39	93.22	65.64
Ours	-	CLIP-ViT	CLIP-Xformer	**74.25**	89.83	93.58	**66.15**

Table 1: Performance comparisons with state-of-the-art methods on CUHK-PEDES dataset. R1, R5, R10 denote the Rank-1, Rank-5, Rank-10 accuracies (%), mAP is the mean average precision (%).

Method	R1	R5	R10	mAP
DSSL(Zhu et al. 2021)	39.05	62.60	73.95	-
SSAN(Ding et al. 2021)	43.50	67.80	77.15	-
LBUL(Wang et al. 2022b)	45.55	68.20	77.85	-
TIPCB (Chen et al. 2022)	46.60	71.70	81.00	36.18
IVT(Shu et al. 2023)	46.70	70.00	78.80	-
ACSA (Ji et al. 2022)	48.40	71.85	81.45	-
CFine(Yan et al. 2022)	50.55	72.50	81.60	-
MCM(Wei et al. 2023)	55.35	77.30	84.25	-
IRRA (Jiang and Ye 2023)	60.25	81.30	88.20	47.52
baseline (CLIP-ViT-B/16)	59.80	81.50	88.30	47.42
Ours	**63.40**	**83.30**	**90.30**	**49.28**

Table 2: Performance comparisons with state-of-the-art methods on RSTPReid dataset.

Method	R1	R5	R10	mAP
MIA (Niu et al. 2020)	46.49	67.14	75.18	-
ViTAA (Wang et al. 2020)	50.98	68.79	75.78	-
SSAN (Ding et al. 2021)	54.23	72.63	79.53	-
TIPCB (Chen et al. 2022)	54.23	72.63	79.53	-
IVT (Shu et al. 2023)	56.04	73.60	80.22	-
CFine (Yan et al. 2022)	60.83	76.55	82.42	-
MCM (Wei et al. 2023)	62.29	77.15	82.52	-
IRRA (Jiang and Ye 2023)	63.46	80.25	85.82	38.06
baseline (CLIP-ViT-B/16)	63.34	80.21	85.73	37.88
Ours	**65.62**	**80.54**	**85.83**	**38.78**

Table 3: Performance comparisons with state-of-the-art methods on ICFG-PEDES dataset.

et al. 2020), and random erasing (Zhong et al. 2020) for image augmentation. The batchsize is set to 64. The maximum length of the textual token sequence is 77. Our model is trained using Adam optimizer (Kingma and Ba 2014) for 60 epochs, with a learning rate initialized at 1×10^{-5} and cosine learning rate decay. The learning rate is gradually increased from 1×10^{-6} to 1×10^{-5} over the 5 warm-up epochs. For the MUM module, both the coupling factor ω and the scale factor s are set to 0.25. The MUM module is only applied during training phase for feature augmentation. During testing phase, we do not use this module. The mask rate of input text token during training phase is set to 30% for CUHK-PEDES and ICFG-PEDES, and 15% for RSTPReid. During the testing phase, the input texts is not masked. The hyper-parameters γ and m in the cm-Circle loss are empirically set to 64 and 0.35. The weight λ_1 of cm-Circle loss is set to 2.0 for ICFG-PEDES and RSTPReid, and 0.25 for CUHK-PEDES. The weight λ_2 of cm-GSR loss is set to 0.5.

Comparison with State-of-the-Art Methods

Results on CUHK-PEDES dataset. As shown in the Table 1, our method surpasses all current state-of-the-art methods, achieving a Rank-1 accuracy of **74.25**% and an mAP of 66.15%. Compared to the methods CFine (Yan et al. 2022) and IRRA (Jiang and Ye 2023), which also employ CLIP pre-trained model as image-text encoders, our method surpasses them by +4.68% and +0.87% in terms of Rank-1, respectively. It is noteworthy that our approach primarily relies on global matching and does not use complex local image-text matching such as (Niu et al. 2020; Yan et al. 2022). Additionally, we also do not leverage external knowledge such as semantic mask (Wang et al. 2020), human pose (Jing et al. 2020), and hierarchical textual parsing (Niu et al. 2020).

Results on RSTPReid dataset. Note that RSTPReid dataset presents complex indoor and outdoor scene variations, making it more challenging. The comparative results on the RSTPReid dataset are shown in Table 2. It is evident that our approach demonstrates a more notable advantage compared

No.	Component			RSTPReid		
	MUM	cm-Circle	cm-GSR	R1	R5	R10
0				59.80	81.50	88.30
1	✓			62.35	82.80	89.05
2		✓		61.50	82.30	88.50
3			✓	61.25	82.20	88.85
4		✓	✓	62.45	82.50	88.95
5	✓	✓		62.80	82.95	89.50
6	✓	✓	✓	**63.40**	**83.30**	**90.30**

Table 4: Ablation study for each proposed component of our method on RSTPReid dataset, No.0 corresponds to baseline.

Method	RSTPReid		
	R1	R5	R10
baseline	59.80	81.50	88.30
baseline+ batch-level UM	61.20	81.85	88.55
baseline+ identity-level UM	61.65	82.40	88.75
baseline+ MUM (multi-granularity)	62.35	82.80	89.05
baseline+ circle loss	60.85	81.85	88.50
baseline+ cm-Circle loss	61.50	82.30	88.50

Table 5: The detailed analysis of the multi-modal uncertainty modeling (MUM) and cross-modal circle loss (cm-Circle).

to other methods. We achieve a Rank-1 accuracy of **63.40**% and an mAP of **49.28**%, significantly surpassing the current SOTA IRRA method by approximately +**3.15**% Rank-1.

Results on ICFG-PEDES dataset. The comparative results on the ICFG-PEDES dataset are presented in Table 3. It is noteworthy that the textual descriptions in the ICFG-PEDES dataset are more focused on individual identities and offer finer granularity. On the ICFG-PEDES dataset, our method still surpasses all existing state-of-the-art methods. We achieve a Rank-1 accuracy of **65.62**%, outperforming the IRRA method by +2.16% in Rank-1 accuracy.

Ablation Study

In this paper, we adopt the CLIP-ViT-B/16 model fine-tuned with the combination of SDM loss $\mathcal{L}_{\text{SDM}}$ and MLM loss $\mathcal{L}_{\text{MLM}}$ as our baseline. The extensive ablation experiments are conducted on top of this baseline to demonstrate the effectiveness of each of our proposed components. Firstly, from the Table 1, 2 and 3, we can observe that our holistic approach consistently yields significant performance improvements over the baseline on three datasets. Compared to the baseline, our method achieves relative improvements of +3.6%, +2.16%, and +1.27% in Rank-1 on RSTPReid, ICFG-PEDES, and CUHK-PEDES, respectively. This validates the effectiveness of our method for TI-ReID.

Effectiveness of the multi-modal uncertainty modeling (MUM). Our MUM module serve as feature augmentation to express richer image-text semantic relationships. The effectiveness of MUM is demonstrated through experimental results involving comparisons between No.0 and No.1, No.2 and No.5, and No.4 and No.6 in the Table 4. For instance, by comparing No.0 and No.1, we observe that solely applying the MUM module leads to a **2.55**% Rank-1 improvement for the baseline on RSTPReid. Furthermore, in the first four rows of Table 5, we experimentally validate the advantage of MUM's coupling of batch-level and identity-level feature variances for multi-granularity uncertainty estimation. We can first see that utilizing either the coarse-grained batch-level uncertainty or fine-grained identity-level uncertainty can enhance the baseline performance. More importantly, when coupling Σ_{batch} and Σ_{ID} to derive multi-granularity uncertainty Σ_{unify} and thus capture more comprehensive and reasonable potential variations, the performance is further improved. This clearly shows the benefits of multi-granularity uncertainty estimation for TI-ReID.

Effectiveness of the cross-modal circle loss (cm-Circle). Our introduced cross-modal circle loss aims to align the global semantic features for positive and negative cross-modal image-text pairs in a self-paced manner. The effectiveness of the cm-Circle loss is demonstrated by comparing results from Table 4 between pair of lines such as No.0 and No.2, No.3 and No.4, and No.1 and No.5. By comparing No.0 and No.2, we can see that optimizing the additional cm-Circle loss results in an **1.7**% Rank-1 improvement to the baseline. We attribute this enhancement primarily to the dynamic adjustment of cross-modal pair weights in the cm-Circle loss and it can enhance the alignment intensity for hard image-text pairs. Furthermore, in the last two rows of Table 5, we compared the cm-Circle loss with conventional circle loss for TI-ReID. We can observe that the cm-Circle loss achieves better performance. This is because cm-Circle loss focuses exclusively on cross-modal pairs and does not optimize negative pairs within text modality. It preserves the intra-modality structure and offers benefits for TI-ReID.

Effectiveness of cross-modal global semantic recovery (cm-GSR). The cm-GSR task is designed to recover the cross-modal semantic of masked *global* text token after the cross-modal interaction, based on the masked language modeling. We verify its effectiveness by conducting comparisons in Table 4 across pairs of rows, including No.0 and No.3, No.2 and No.4, and No.5 and No.6. As we can see, incorporating the cm-GSR task alone results in a Rank-1 improvement of 1.45% over baseline. In addition, applying it on top of the MUM module and cm-Circle loss further amplifies semantic alignment capability, resulting in better performance. These results confirm the necessity of cm-GSR task and its potential on promoting a more comprehensive image-text semantic alignment for TI-ReID.

Conclusion

This paper presents a novel method that unifies multi-modal uncertainty modeling and semantic alignment for text-to-image person Re-ID. We explicitly model the uncertainty in pedestrian images and textual descriptions, using Gaussian distributions to depict image/text features and estimates multi-granularity uncertainty by jointly using batch-level and identity-level variances. We further propose bi-directional cross-modal circle loss to more effectively align probabilistic image and text features. Moreover, we develop cm-GSR task to promote more comprehensive image-text alignment. Extensive experiments on TI-ReID benchmarks show the effectiveness and superiority of our method.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grant No. 62272430, Grant No. 62121002).

References

Chang, J.; Lan, Z.; Cheng, C.; and Wei, Y. 2020. Data uncertainty learning in face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5710–5719.

Chen, Y.; Zhang, G.; Lu, Y.; Wang, Z.; and Zheng, Y. 2022. TIPCB: A simple but effective part-based convolutional baseline for text-based person search. *Neurocomputing*, 494: 171–181.

Chun, S.; Oh, S. J.; De Rezende, R. S.; Kalantidis, Y.; and Larlus, D. 2021. Probabilistic embeddings for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8415–8424.

Cubuk, E. D.; Zoph, B.; Shlens, J.; and Le, Q. V. 2020. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 702–703.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ding, Z.; Ding, C.; Shao, Z.; and Tao, D. 2021. Semantically self-aligned network for text-to-image part-aware person reidentification. *arXiv preprint arXiv:2107.12666*.

Farooq, A.; Awais, M.; Kittler, J.; and Khalid, S. S. 2022. AXM-Net: Implicit Cross-Modal Feature Alignment for Person Re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 4477–4485.

Gao, C.; Cai, G.; Jiang, X.; Zheng, F.; Zhang, J.; Gong, Y.; Peng, P.; Guo, X.; and Sun, X. 2021. Contextual non-local alignment over full-scale representation for text-based person search. *arXiv preprint arXiv:2101.03036*.

Han, X.; He, S.; Zhang, L.; and Xiang, T. 2021. Text-Based Person Search with Limited Data. In *BMVC*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Ji, Y.; Wang, J.; Gong, Y.; Zhang, L.; Zhu, Y.; Wang, H.; Zhang, J.; Sakai, T.; and Yang, Y. 2023. MAP: Multimodal Uncertainty-Aware Vision-Language Pre-Training Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23262–23271.

Ji, Z.; Hu, J.; Liu, D.; Wu, L. Y.; and Zhao, Y. 2022. Asymmetric cross-scale alignment for text-based person search. *IEEE Transactions on Multimedia*.

Jiang, D.; and Ye, M. 2023. Cross-Modal Implicit Relation Reasoning and Aligning for Text-to-Image Person Retrieval. In *CVPR*.

Jing, Y.; Si, C.; Wang, J.; Wang, W.; Wang, L.; and Tan, T. 2020. Pose-guided multi-granularity attention network for text-based person search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 11189–11196.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Li, S.; Xiao, T.; Li, H.; Zhou, B.; Yue, D.; and Wang, X. 2017. Person search with natural language description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1970–1979.

Li, X.; Dai, Y.; Ge, Y.; Liu, J.; Shan, Y.; and Duan, L.-Y. 2022. Uncertainty modeling for out-of-distribution generalization. *arXiv preprint arXiv:2202.03958*.

Niu, K.; Huang, Y.; Ouyang, W.; and Wang, L. 2020. Improving description-based person re-identification by multigranularity image-text alignments. *IEEE Transactions on Image Processing*, 29: 5542–5556.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Sarafianos, N.; Xu, X.; and Kakadiaris, I. A. 2019. Adversarial representation learning for text-to-image matching. In *Proceedings of the IEEE/CVF international conference on computer vision*, 5814–5824.

Shao, Z.; Zhang, X.; Fang, M.; Lin, Z.; Wang, J.; and Ding, C. 2022. Learning Granularity-Unified Representations for Text-to-Image Person Re-identification. In *Proceedings of the 30th ACM International Conference on Multimedia*, 5566–5574.

Shu, X.; Wen, W.; Wu, H.; Chen, K.; Song, Y.; Qiao, R.; Ren, B.; and Wang, X. 2023. See finer, see more: Implicit modality alignment for text-based person retrieval. In *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part V*, 624–641. Springer.

Sun, Y.; Cheng, C.; Zhang, Y.; Zhang, C.; Zheng, L.; Wang, Z.; and Wei, Y. 2020. Circle loss: A unified perspective of pair similarity optimization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6398–6407.

Wang, Z.; Fang, Z.; Wang, J.; and Yang, Y. 2020. Vitaa: Visual-textual attributes alignment in person search by natural language. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, 402–420. Springer.

Wang, Z.; Zhu, A.; Xue, J.; Wan, X.; Liu, C.; Wang, T.; and Li, Y. 2022a. CAIBC: Capturing All-round Information Beyond Color for Text-based Person Retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, 5314–5322.

Wang, Z.; Zhu, A.; Xue, J.; Wan, X.; Liu, C.; Wang, T.; and Li, Y. 2022b. Look Before You Leap: Improving Text-based Person Retrieval by Learning A Consistent Cross-modal Common Manifold. In *Proceedings of the 30th ACM International Conference on Multimedia*, 1984–1992.

Wei, D.; Zhang, S.; Yang, T.; and Liu, J. 2023. Calibrating Cross-modal Feature for Text-Based Person Searching. *arXiv preprint arXiv:2304.02278*.

Wu, Y.; Yan, Z.; Han, X.; Li, G.; Zou, C.; and Cui, S. 2021. LapsCore: language-guided person search via color reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1624–1633.

Yan, S.; Dong, N.; Zhang, L.; and Tang, J. 2022. CLIP-Driven Fine-grained Text-Image Person Re-identification. *arXiv preprint arXiv:2210.10276*.

Ye, M.; Shen, J.; Lin, G.; Xiang, T.; Shao, L.; and Hoi, S. C. H. 2022. Deep Learning for Person Re-Identification: A Survey and Outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6): 2872–2893.

Yu, T.; Li, D.; Yang, Y.; Hospedales, T. M.; and Xiang, T. 2019. Robust person re-identification by modelling feature uncertainty. In *Proceedings of the IEEE/CVF international conference on computer vision*, 552–561.

Zhang, Y.; and Lu, H. 2018. Deep cross-modal projection learning for image-text matching. In *Proceedings of the European conference on computer vision (ECCV)*, 686–701.

Zheng, Z.; Zheng, L.; Garrett, M.; Yang, Y.; Xu, M.; and Shen, Y.-D. 2020. Dual-path convolutional image-text embeddings with instance loss. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 16(2): 1–23.

Zhong, Z.; Zheng, L.; Kang, G.; Li, S.; and Yang, Y. 2020. Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 13001–13008.

Zhu, A.; Wang, Z.; Li, Y.; Wan, X.; Jin, J.; Wang, T.; Hu, F.; and Hua, G. 2021. DSSL: deep surroundings-person separation learning for text-based person retrieval. In *Proceedings of the 29th ACM International Conference on Multimedia*, 209–217.