

RESEARCH

Open Access



Uncertainty-aware coarse-to-fine alignment for text-image person retrieval

Yifei Deng¹, Zhengyu Chen², Chenglong Li^{2*} and Jin Tang¹

Abstract

Text-to-image person retrieval, a fine-grained cross-modal retrieval problem, aims to search for person images from an image library that match a given textual caption. Existing text-to-image person retrieval methods usually use fixed-point embedding to express the semantics of the two modalities and perform multi-granularity alignment between modalities in the embedding space. However, owing to the inherent mutual one-to-many correspondence between images and texts, it is often difficult for fixed-point embedding methods to adequately capture this relationship, leading to erroneous retrieval results. To address this problem, we propose a novel uncertainty-aware coarse-to-fine alignment method, which first maps fixed-point embedding to probability distributions and then aligns two modalities in terms of distributions and sampling points at a coarse-to-fine granularity, for accurate text-to-image person retrieval. Specifically, we first introduce two contrastive learning tasks of distribution contrast learning and point contrast learning, to achieve coarse-grained inter-modal alignment with uncertainty-aware. The distribution contrast learning task ensures that distributions with the same identity are as similar as possible across modalities through distribution-based contrastive learning. The point contrast learning task performs the contrastive learning of inter-modal and intra-modal sampling points, which not only models rich and diverse cross-modal associations, but also optimizes the learning of distributions. For the fine-grained association requirements of text-to-image person retrieval, we design the task of uncertainty-aware attribute masking language reconstruction, which achieves fine-grained alignment by randomly masking attribute words in the text and reconstructing them via inter-modal sample point interactions. Extensive experiments on two public datasets demonstrate the superior performance of our method.

Keywords: Cross-modal retrieval, Uncertainty-aware, Coarse-to-fine alignment, Probability distribution, Contrastive learning

1 Introduction

Text-to-image person retrieval is a retrieval task within multi-modal learning [1–3], as well as an extended direction of person re-identification (Re-ID) [4]. Its goal is to efficiently retrieve the corresponding target person images from a large-scale image database based on the given textual descriptions. Unlike previous image-based person Re-

ID [5], it leverages flexible and easily accessible textual descriptions for retrieval [6], enabling wide applications in scenarios, in which image information is not readily available. This task requires the model to bridge the modal gap between the text and the image. Additionally, compared with traditional text-to-image retrieval tasks, this task focuses specifically on retrieving images of a person. Given the often subtle differences individuals, the model needs to improve its fine-grained alignment capabilities to more accurately identify and distinguish between different persons.

The text-to-image person retrieval task faces three main challenges: modality differences, cross-modal semantic

*Correspondence: lcl1314@foxmail.com

²Information Materials and Intelligent Sensing Laboratory of Anhui Province, Anhui Provincial Key Laboratory of Security Artificial Intelligence, School of Artificial Intelligence, Anhui University, Hefei, 230601, China
Full list of author information is available at the end of the article

gaps, and semantic alignment issues. Text is often concise and abstract, whereas images are concrete and intuitive, making matching between the two complex. Additionally, differences in semantic expression between text and images create cross-modal semantic gaps. Accurately aligning the semantics between text and image, particularly during fine-grained feature matching, remains a key to improving retrieval accuracy.

The key to text-to-image person retrieval lies in establishing an accurate association between the two modalities. To address this challenge, researchers have proposed various methods. For instance, some studies introduce cross-modal matching loss functions [7], which learn global alignment by separately extracting features from text and images and then calculating their similarity. However, these methods fail to capture the fine-grained characteristics adequately. To address this limitation, researchers have explored local matching approaches. Some efforts attempt to explicitly match body parts with words in the text [8] or implicitly achieve alignment through attention mechanisms [9]. Other studies leverage linguistic toolkits [10] to decompose the text into phrases for alignment [6, 11]. However, these methods often have high computational costs and overly rely on the reliability of the toolkits, which can weaken the connection between nouns and phrases and degrade the quality of text features. In contrast, Jiang and Ye [12] successfully achieved implicit local alignment by introducing a masked language modeling task. This method encourages the model to learn cross-modal fine-grained relationships, significantly improving retrieval performance.

The aforementioned methods have advanced the field of text-to-image person retrieval, but they rely on fixed-point embeddings to represent multi-modal semantic information. This approach struggles to effectively model the inherent uncertainty within each modality, potentially resulting in incorrect associations between images and text, which can adversely impact retrieval performance. Specifically, fixed-point embeddings map text and image data into a common high-dimensional embedding space, in which each text or image is represented as a fixed point. However, this method results in significant limitations when dealing with complex one-to-many and many-to-one correspondences between text and images. For instance, in commonly used text-image person retrieval datasets, such as ICFG-PEDES [13] and RSTPReid [14], a single description can match multiple images of a person, and a single image can fit several different descriptions, as shown in Fig. 1. Such complex correspondences make it difficult for fixed-point embeddings to capture this diversity, leading to an oversight of polysemy in language. This problem reduces the accuracy and robustness of cross-modal alignment, thereby increasing the likelihood of retrieval errors.

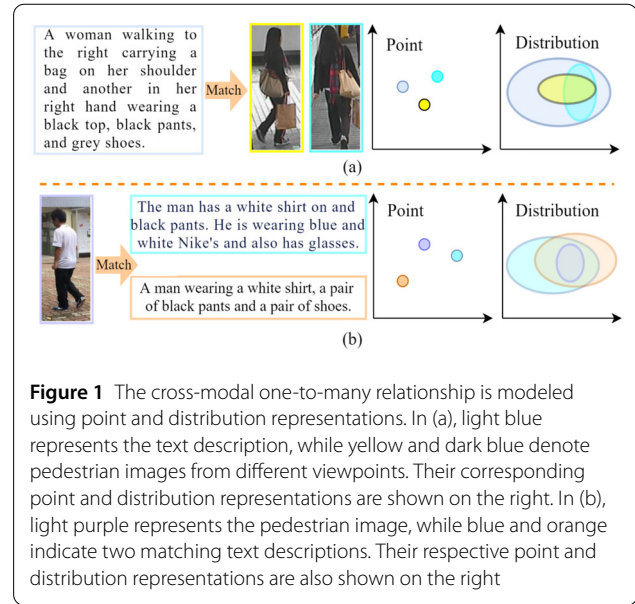


Figure 1 The cross-modal one-to-many relationship is modeled using point and distribution representations. In (a), light blue represents the text description, while yellow and dark blue denote pedestrian images from different viewpoints. Their corresponding point and distribution representations are shown on the right. In (b), light purple represents the pedestrian image, while blue and orange indicate two matching text descriptions. Their respective point and distribution representations are also shown on the right

To address the above problem, we propose a novel uncertainty-aware coarse-to-fine alignment network (UANet). It addresses the aforementioned issue through uncertainty-aware methods. Specifically, in the text-to-image cross-modal person retrieval task, uncertainty-aware refers to expanding fixed-point embedding-based feature representations into distributional representations to capture and convey rich semantic details. Traditional fixed-point embeddings map each sample to a deterministic feature vector, overlooking the natural variability and uncertainty in sample features across different modalities and viewpoints. Through uncertainty-aware methods, each sample embedding can be mapped to a probability distribution, such as a Gaussian distribution, which extends the sample's feature space into a more inclusive distributional representation. This effectively reflects the inherent diversity and ambiguity of the sample. In cross-modal person retrieval, this uncertainty-aware representation is crucial for semantic consistency. Given the inherent differences in modal characteristics between text descriptions and image samples, such as the abstract nature of text and the visual nature of images, uncertainty-aware can effectively capture the similarities and differences between modality features through distributional parameters (e.g., mean and variance).

UANet maps fixed-point embedding to probability distribution and aligns the two modalities in terms of both distributions and sampling points from coarse to fine, resulting in diverse and diverse representations of embedded space. Specifically, to map fixed-point embedding into a Gaussian distribution representation, we first design a distribution generation module (DGM), which enhances intra-modal interactions over self-attention, capturing rich information and resulting in accurate predictions

of the mean and variance. Then, to model the complex one-to-many correspondence, we perform uncertainty-aware alignment from coarse to fine grained from both distribution and sampling point perspectives. Specifically, to realize uncertainty-aware coarse-grained alignment, distribution contrastive learning (DCL) and point contrastive learning (PCL) are proposed. DCL ensures that the distributions of the same IDs are similar as possible as across modalities by comparing the distributional differences. PCL builds rich cross-modal associations by contrastive learning of the sampling points and the mean points. To achieve more accurate fine-grained alignment, the uncertainty-aware attribute masked language reconstruction (U-AMLR) task is designed, which reduces interference from task-irrelevant vocabulary by masking person attributes. Moreover, it utilizes the interaction between the mean and the sampling points to reconstruct the masked words, leading to a more precise fine-grained associations between modalities.

Overall, the main contributions of this paper can be summarized as follows:

- 1) We propose uncertainty-aware coarse-to-fine alignment network called UANet, which effectively models one-to-many correspondence between modalities by transforming fixed-point embeddings into Gaussian probability distribution representations.
- 2) We design two tasks, distributional contrastive learning called DCL and point contrastive learning called PCL, to build rich and diverse embedding space representations.
- 3) We design uncertainty-aware attribute masked language reconstruction task, termed as U-AMLR, to further improve the accuracy of fine-grained alignment.
- 4) Extensive experiments on publicly available text-to-image person retrieval datasets demonstrate the superiority of our approach over state-of-the-art methods.

2 Related work

2.1 Text-to-image person retrieval

With the rapid advancement of deep learning [15], text-to-image person retrieval has gained considerable attention. It was first introduced by Li et al. [16], aiming to establish fine-grained associations between two heterogeneous modalities within a shared embedding space. In terms of feature extraction, early methods utilized separately pre-trained encoders, such as visual geometry group (VGG) for images and long short-term memory (LSTM) for text, to extract features [6, 17]. Later, more advanced encoders, such as residual network (ResNet) and bi-directional encoder representations from transformers (BERT) were adopted to extract features from images and text, respectively [7, 18]. However, these encoders typically handle

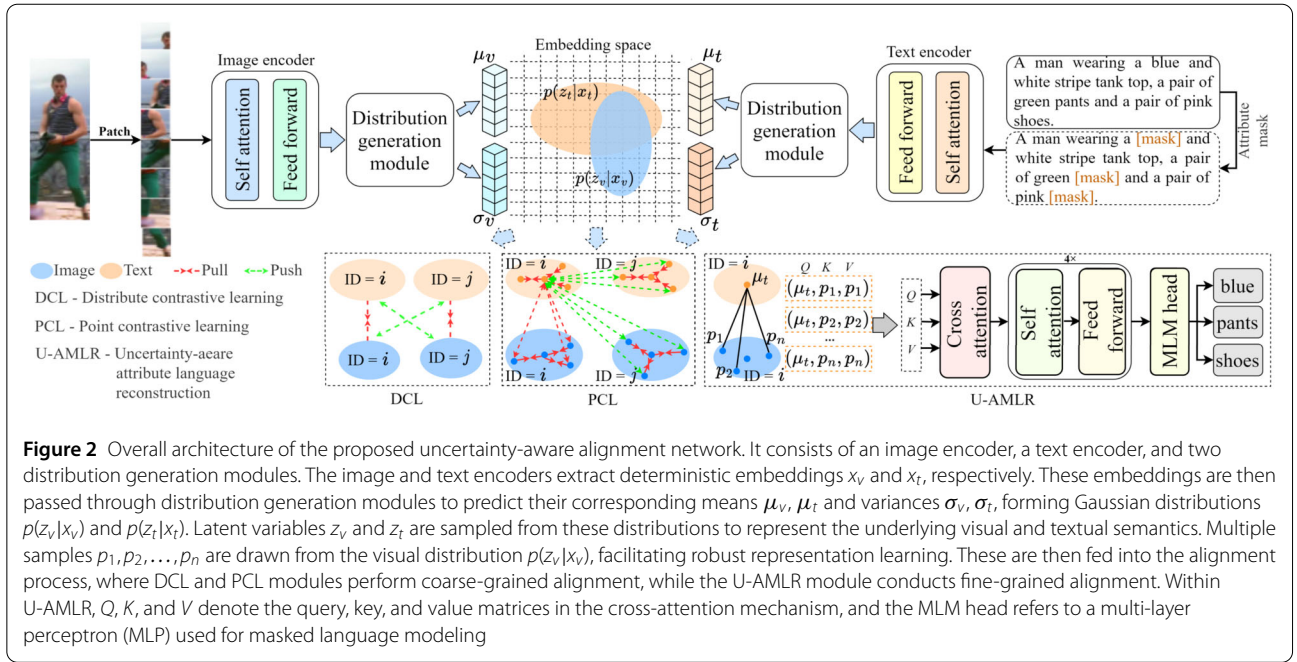
data from a single modality and lack cross-modal capabilities. To address this, recent studies [12, 19] have explored the potential of contrastive language-image pre-training (CLIP) encoders for feature extraction and applied them to person retrieval tasks, significantly enhancing retrieval performance.

In terms of alignment improvements, early studies applied matching losses on extracted image and text features to learn alignment capabilities [20]. However, these methods were limited to coarse-grained global matching and failed to capture the fine-grained characteristics of the task. To address this issue, later works focused on enhancing local feature learning, such as by explicitly segmenting body parts [8], dividing text into phrases [11], and adding attribute annotations [21], to improve local matching. In contrast, recent studies [12] have achieved implicit local alignment by introducing masked language modeling tasks, enabling the model to learn cross-modal fine-grained relationships and significantly enhancing retrieval performance. While these methods have advanced the field of text-to-image person retrieval, they still rely on fixed-point embeddings to represent the semantics of different modalities, limiting the ability of model to handle complex one-to-many and many-to-many correspondences between text and images.

2.2 Probability distribution embedding

Probabilistic embedding in deep learning is a method that combines probabilistic theory with embedding techniques, primarily used to handle data with uncertainty and probabilistic characteristics. In the context of deep learning, probabilistic embedding is often employed to build models capable of processing uncertainties and noises, thereby accurately describing and predicting complex phenomena in the real world.

Previously, in the field of natural language processing, some works [22] utilized multi-variate Gaussian distributions to represent words, aiming to capture the asymmetric relationships among them. In computer vision, other works introduced Gaussian representations for specific tasks, to model visual uncertainties, such as face recognition [23] and person Re-ID [24]. Within cross-modal retrieval, the need arises to construct a common representation space for samples from multiple modalities (vision and language). However, the diversity in the correspondence between images and their descriptions makes this task particularly challenging, and deterministic functions are insufficient to capture such one-to-many relationships. Consequently, recent researchers [25] have introduced probabilistic embedding into multi-modal retrieval tasks. Nevertheless, compared with general cross-modal retrieval, the introduction of probabilistic embedding in text-to-image person retrieval is still in its infancy. Therefore, we propose applying probabilistic embeddings



to the text-to-image person retrieval model to account for the inherent uncertainty within each modality and to better model complex relationships, such as one-to-many matching.

3 Method

In this section, we first overview the proposed uncertainty-aware coarse-to-fine alignment network called UANet, and then present the details of UANet and the testing method.

3.1 Overview

The overall structure of UANet is shown in Fig. 2. Specifically, each input pedestrian image is first divided into N non-overlapping patches, where $N = (HW)/(P_h P_w)$, with H and W representing the height and width of the image, and P_h and P_w representing the height and width of each patch. The image encoder is then used to extract the fixed-point embedding of the visual patches, denoted as x_v . Next, the DGM uses these embeddings to predict the mean μ_v and variance σ_v , and maps the embedding to a probability distribution $p(z_v|x_v)$, where z_v is the latent variable sampled from the conditional distribution $p(z_v|x_v)$, representing the latent representation of x_v . For each text description, random attribute masking is performed first, and the processed text is fed into the text encoder to obtain the fixed-point embedding x_t . Similar to image processing, the text embedding is also mapped to a probability distribution $p(z_t|x_t)$ after passing through the DGM, where z_t is the latent variable sampled from the conditional distribution $p(z_t|x_t)$, representing the latent representation of

the text description x_t . Next, UANet combines both distribution and sampling point perspectives for coarse to fine alignment. Specifically, for z_v and z_t , we first perform the DCL task to drive the pair-wise distributions to be as similar as possible. Moreover, the PCL task establishes complex associations between two modalities from the perspective of sampling points. Consequently, with DCL and PCL, UANet achieves inter-modal uncertainty-aware coarse-grained alignment. Moreover, we perform the U-AMLR task on these sampling points.

3.2 Backbone

To fully utilize the current knowledge of multi-modal pre-trained large models, we directly utilize the image and text encoders in CLIP [26] to extract the embedding for each modality.

Image encoder. Since the image encoder of CLIP utilizes the ViT [27] architecture, for a given person image, it is first divided into N non-overlapping and same-sized image patches. After linear projection, these patches are mapped into 1D tokens. To represent the global information of image and preserve the positional information, these tokens are additionally appended with the classification token ([CLS]) and the corresponding positional encoding. These token sequences are fed into the Transformer to model the relationship between patches and patches.

Text encoder. For the text description, it is first tokenized. In other words, the corresponding words are transformed into the corresponding encodings. Then, the start of sequence token ([SOS]) and the end of sequence token ([EOS]) are added at the beginning and end positions of

text description, which is consistent with the tokenization of text in CLIP. Afterwards, the tokenized text sequences are fed into the encoder based on Transformer, which models the associations between each token by masking self-attention. Eventually, the [EOS] tokens that are used to characterize the global information of text are linearly projected into the embedding space of the union to obtain a global representation of text.

3.3 Distribution generation

To model the uncertainty in both the text and image modalities, we propose DGM to map the point embeddings of different modalities to multivariate Gaussian distributions. Specifically, DGM takes the fixed point embedding of a modality as input and outputs the mean vector and variance vector. For example, in the visual modality, given any fixed point embedding x_v , DGM defines its representation z_v as a Gaussian distribution in the embedding space:

$$p(z_v|x_v) = N_v(z_v; \mu_v, \sigma_v^2), \quad (1)$$

where N_v represents the diagonal multi-variate Gaussian distribution, μ_v represents the mean vector of distribution of x_v , σ_v^2 denotes the variance vector. In particular, since the mean vector represents the center of the distribution, its semantics should be closest to the main semantics of the image. Therefore, we directly use the fixed point embedding x_v of the image as the mean vector of the distribution. On the other hand, the variance of the distribution captures the uncertainty and semantic variation in the data. To obtain it, we extend the fixed point embedding. Specifically, we first enhance the fixed point embedding x_v using a self-attention mechanism. Then, we use residual connections to ensure the efficient transmission of the enhanced features. Finally, a fully connected layer is used to map these enhanced features to a new embedding space to generate the variance. The computation process is as follows:

$$\mu_v = x_v, \quad (2)$$

$$\sigma_v^2 = \text{FC}(\text{SA}(x_v) + F(\text{SA}(x_v))), \quad (3)$$

where SA represents multi-head self-attention, F is the feedforward network, and FC is the fully connected layer, with x_v being the fixed point embedding of the image. By adopting this method, we are able to obtain both the mean and variance for the text and image modalities simultaneously. With DGM, each sample from the visual or textual modality is mapped to a Gaussian distribution in the high-dimensional embedding space.

3.4 Uncertainty-aware coarse-grained alignment

To achieve uncertainty-aware coarse-grained alignment between modalities, we design DCL and PCL, which jointly construct rich multi-modal embedding representations from different perspectives.

Distribution contrastive learning. In DCL, to measure the similarity between two distributions, we first calculate their 2-Wasserstein distance. Compared with the Kullback-Leibler (KL) divergence and Jensen-Shannon (JS) divergence, the 2-Wasserstein distance can handle the case of non-overlapping distributions and provide information about the gradient that can help in optimization. Since all covariance matrices are diagonal matrices in our modeled distribution, and the 2-Wasserstein distance can be simplified as

$$D_{2w} = \|\mu_v - \mu_t\|_2^2 + \|\sigma_v - \sigma_t\|_2^2, \quad (4)$$

where μ_v and μ_t represent the means of the image and text embeddings after being mapped to Gaussian distributions, and σ_v and σ_t represent the standard deviations of the image and text embeddings after being mapped to Gaussian distributions. Through the use of the 2-Wasserstein distance, we can quantify the similarity between two multivariate Gaussian distributions as

$$s(v, t) = -a \times D_{2w} + b, \quad (5)$$

where $s(v, t)$ measures the similarity between visual and textual two distributions, a is the proportionality coefficient and b is a shift value. Next, the probability p_{ij} of a text-to-image matching pair can be calculated as

$$p_{ij} = \frac{\exp(s(v_i, t_j)/\tau_1)}{\sum_{k=1}^N \exp(s(v_i, t_k)/\tau_1)}, \quad (6)$$

where τ_1 is a temperature coefficient that is detailed in Ref. [12] to enable model to focus on hard-negative samples. It can increase the correlation between matched image-to-text pairs and the variance between unmatched text-to-image pairs. Finally, following CMPM [7], the DCL loss from image to text in a batch can be obtained as

$$L_{D-2t} = L_{KL}(p_i||q_i) = \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^N p_{ij} \log\left(\frac{p_{ij}}{q_{ij} + \epsilon}\right), \quad (7)$$

where L_{KL} represents Kullback-Leibler divergence, p_i denotes the predicted matching distribution between the i -th image and all text, while q_i represents the corresponding ground truth matching distribution, $q_{ij} = y_{ij} / \sum_{k=1}^N y_{ik}$ is the true matching probability, y_{ik} indicates whether the i -th image matches the k -th text, where $y_{ik} = 1$ if they match, and $y_{ik} = 0$ otherwise. ϵ is the number to avoid numerical

problems. Similarly, the DCL loss L_{D-t2i} from text to image is similar to above process. By combining these two directions losses, the final DCL loss is formed as

$$L_{DCL} = L_{D-i2t} + L_{D-t2i}. \quad (8)$$

Point contrastive learning. Since the learning of distributions lacks explicit constraints, aligning two modalities from the perspective of distribution alone is not sufficient. To this end, PCL is proposed to perform inter-modal and intra-modal contrastive learning from the perspective of the sampling points. It not only establishes accurate and rich cross-modal associations, but also improves the learning of distributions.

To build diverse cross-modal associations, we perform cross-modal point contrastive learning between the sampled points from the distribution of one modality and the distribution mean of the other modality. Specifically, for the k -th sampled point v_{ik} from the distribution of the i -th image v_i , and the distribution mean μ_t^j of the j -th text t_j , we first measure their similarity using cosine similarity. Consistent with the DCL method, we introduce a temperature coefficient τ_2 to adjust the focus on hard negative samples. Then, the matching probability $p_{ik,j}$ between v_{ik} and μ_t^j is computed as follows:

$$p_{ik,j} = \frac{\exp((v_{ik} \times \mu_t^j)/\tau_2)}{\sum_{j=1}^N \exp((v_{ik} \times \mu_t^j)/\tau_2)}, \quad (9)$$

thus, the contrastive learning loss for a single sample point k , denoted as $L_{k_{i2t}}$, can be expressed as

$$L_{k_{i2t}} = \text{KL}(p_{ik,j} || q_{i,j}) = \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^N p_{ik,j} \log\left(\frac{p_{ik,j}}{q_{i,j} + \epsilon}\right), \quad (10)$$

where τ_2 is a temperature coefficient that enables model to focus on hard-negative samples, $q_{i,j}$ is the true matching probability, N represents the number of samples in a batch, and ϵ is a small constant (set to 1.0×10^{-8}) to prevent division by zero errors.

Thus the total sample point loss L_{P-i2t} from image to text can be calculated in the following way:

$$L_{P-i2t} = \frac{1}{S} \sum_{k=1}^S L_{k_{i2t}}, \quad (11)$$

where S is the number of sample points. Similarly, we can compute L_{P-t2i} . The final inter-modal point-contrast learning loss can be expressed as

$$L_{\text{Inter-P}} = L_{P-i2t} + L_{P-t2i}. \quad (12)$$

Moreover, we believe that if the mapped probability distribution is accurate, any of these sampling points can be considered as a positive sample of the mean of the current distribution, whereas the means of other distributions are considered negative samples. Therefore, we introduce intra-modal point contrastive learning to constrain the similarity of sample and mean points within a single distribution, which can improve the learning of the distribution. Specifically, taking the k -th point v_{ik} sampled from the distribution mapped by image i as an example, it is computed as

$$L_{k_{v-s2m}} = -\frac{1}{N} \sum_{i=1}^N \log \frac{\exp((v_{ik} \times \mu_v^j)/\tau_2)}{\sum_{j=1}^N \exp((v_{ik} \times \mu_v^j)/\tau_2)}, \quad (13)$$

where μ_v^j represents the mean of the j -th image after being mapped to the Gaussian probability distribution, $L_{k_{v-s2m}}$ denotes the image-modal contrast learning loss at the k -th sampling point.

$$L_{v-s2m} = \frac{1}{S} \sum_{k=1}^S L_{k_{v-s2m}}, \quad (14)$$

where L_{v-s2m} denotes the intra-modal contrast learning loss of image. Similarly, we can compute L_{t-s2m} . Therefore, the intra-modal point-contrastive learning loss can be expressed as

$$L_{\text{Intra-P}} = L_{v-s2m} + L_{t-s2m}. \quad (15)$$

Finally, the PCL loss can be expressed as

$$L_{\text{PCL}} = L_{\text{Inter-P}} + L_{\text{Intra-P}}. \quad (16)$$

3.5 Uncertainty-aware fine-grained alignment

DCL and PCL align the two modalities from the distribution and sampling point perspectives, achieving coarse-grained alignment across modal uncertainty perceptions. However, relying only on DCL and PCL makes it difficult to establish fine-grained associations between text entities and human body details.

To realize accurate fine-grained alignment between modalities, we design the U-AMLR task. Compared with the masked language modeling (MLM), U-AMLR focuses more on task-relevant words reconstruction and reduces the influence of task-irrelevant words (e.g., prepositions, articles, etc.). With the help of uncertainty-aware reconstruction strategies, U-AMLR is able to achieve more accurate reconstruction than fixed-point embedding. Specifically, based on the attribute vocabulary of statistics [8], we design an attribute-based masked strategy. Initially, we assign a random masking probability to each word in the text

description. Words with a masking probability below 15% are masked at a 75% chance if they are presented in the attribute library. Otherwise, they are masked with a 25% probability.

To achieve more precise recovery of masking words, we employ an uncertainty-aware strategy in which samples multiple points from the image distribution to aid in the reconstruction. Specifically, we use multi-head cross-attention mechanism to fuse the image and the attribute mask text representations. In this process, the mean distribution of attribute mask text is used as q , and the sampling points drawn from the image distribution are used as k and v . Considering the i -th sampling point p_i as an example, the interaction with the mean of the text distribution can be expressed as

$$\text{MCA}(q_{\mu_t}, k_{p_i}, v_{p_i}) = \text{Softmax}\left(\frac{q_{\mu_t} k_{p_i}^T}{\sqrt{d}}\right) v_{p_i}, \quad (17)$$

where q_{μ_t} denotes that the mean of the distribution mapped by text t is treated as a query, k_{p_i} and v_{p_i} denote that the embedding of the i -th sample point of the distribution of image is treated as key and value, and d denotes the embedding dimension, MCA denotes multi-head cross attention. Following the interaction, we enhance the resulting embedding via a self-attention (SA) mechanism. Finally, a multi-layer perceptron (MLP) is employed to predict the masked words:

$$\text{Rec}_i = \text{MLP}(\text{SA}(\text{MCA}(q_{\mu_t}, k_{p_i}, v_{p_i}))), \quad (18)$$

where Rec_i denotes the reconstruction result using the i -th sampling point.

For the reconstructed words, we calculate the cross-entropy (CE) loss between them and the ground truth (GT) as follows:

$$L_{\text{rec}} = \sum_{i=1}^{S+1} \text{CE}(\text{Rec}_i, \text{GT}), \quad (19)$$

where L_{rec} represents the loss of reconstructed words, $S+1$ denotes S sampling points and a mean point vector.

3.6 Sampling points

It is important to emphasize that sampling points directly from the distribution can hinder the back propagation of the gradient during training. To address this issue, we employ the reparameterization technique mentioned in Ref. [28]. Specifically, we take a random noise ϵ from the standard normal distribution and reparameterize it as

$$z = \mu + \epsilon\sigma, \quad (20)$$

where μ and σ represent mean and standard deviations of distribution, respectively. In this way, the output z obeys the predictive distribution from DGM and is trainable.

Moreover, if we only rely on the above loss function without constraining the variance of distribution, all sampling points tend to cluster at the optimal location, resulting in the variance shrinking to 0. This will degrade the probability distribution into a single-point distribution, which in turn loses the ability to characterize uncertainty. To avoid this situation, we introduce a regularized loss L_{reg} in Ref. [29] to ensure that the model is able to effectively learn about uncertainty.

3.7 Test-time variants

While distribution embedding models address modal uncertainty and tackle the one-to-many mapping issue, solely relying on distribution similarity to determine the matching probability might be inadequate because of the absence of explicit supervision in learning the distributions. Fixed-point embedding, on the other hand, usually provide more accurate semantic information despite ignoring uncertainty. Therefore, each of them has its own advantages, and the retrieval accuracy can be further improved by combining the distribution similarity and the similarity of fixed points. Specifically, we design a weighted fusion strategy to accurately combine distributional similarity and the mean point similarity, which can be denoted as

$$\text{Simi}(I, T) = \alpha D_{\text{simi}}(I, T) + (1 - \alpha) P_{\text{simi}}(I, T). \quad (21)$$

Here, $\text{Simi}(I, T)$ represents the similarity between image I and text T , α is a hyperparameter set to 0.7, and D_{simi} and P_{simi} representing the distribution similarity and mean point similarity, respectively. D_{simi} is derived via Eqs. (4) and (5), while P_{simi} is computed by calculating the cosine similarity between means.

4 Experiment

4.1 Datasets and evaluation metrics

We carry out extensive experiments on two publicly available datasets RSTPReid [14] and ICFG-PEDES [13].

The RSTPReid dataset is an extension of the MSMT17 dataset [30] specifically designed for text-based person ReID tasks in real-world scenarios, and compiles a vast collection of image resources from 15 different surveillance cameras. This dataset encompasses a total of 20,505 high-resolution images belonging to 4101 unique pedestrians, with each pedestrian having five photographs captured at different time points and angles, comprehensively covering intricate indoor and outdoor environmental variations along with diverse background scenarios.

The ICFG-PEDES dataset is specifically designed for the task of text-based person retrieval. It compiles a total of

Table 1 Comparison results on the ICFG-PEDES [13] and RSTPReid [14] datasets. The best and second best results are bolded and underlined, respectively

Method	ICFG-PEDES			RSTPReid			Params (Mb)	Inference time (ms)
	Rank-1	Rank-5	Rank-10	Rank-1	Rank-5	Rank-10		
CMPM [7]	43.51	65.44	74.26	-	-	-	-	-
ViTAA [8]	50.98	68.79	75.78	-	-	-	-	-
SSAN [13]	54.23	72.63	79.53	43.50	67.80	77.05	98	1.01
IVT [31]	56.04	73.60	80.22	46.70	70.00	78.80	-	-
CFine [19]	60.83	76.55	82.42	50.55	72.50	81.60	-	-
SRCF [32]	57.18	75.01	81.49	-	-	-	199	6.90
LGUR [33]	57.42	74.97	81.45	-	-	-	148	15.40
BEAT [34]	58.25	75.92	81.96	48.10	73.10	81.30	-	-
LCR ² S [35]	57.93	76.08	82.40	54.95	76.65	84.70	-	-
MANet [9]	59.44	76.80	82.75	-	-	-	-	-
UniPT [36]	60.09	76.19	82.46	51.85	74.85	82.85	-	-
IRRA [12]	<u>63.46</u>	<u>80.25</u>	85.82	<u>60.20</u>	<u>81.30</u>	<u>88.20</u>	191	6.40
CSKT [37]	58.90	77.31	83.56	57.75	81.30	88.35	-	-
CAPL [38]	56.61	76.29	82.86	-	-	-	-	-
Ours	63.97	80.27	<u>85.54</u>	61.85	81.85	89.15	207	7.00

54,522 images from 4102 unique identities across multiple diverse scenarios, with each image accompanied by a unique and detailed textual description. Compared with the previous two datasets, these descriptions average 37.2 words in length, resulting in a more specific portrayal of details.

We use the commonly used Rank-1, Rank-5, and Rank-10 metrics [12] in this task to evaluate the performance of UANet.

4.2 Implementation details

For the text and image input to the model, they are first simply pre-processed. The text is tokenized and the maximum length is set to 77. The image is scaled to 384×128 and a series of data enhancement operations, such as scaling, random horizontal flipping, random cropping and random erasing, are performed. For feature extraction, we use the large-scale pre-training model CLIP as visual and text encoders and initialize these encoders with pre-training parameters. In DCL, the value of τ_1 is set to 0.03, and b is a learnable parameter that is initialized to 6. In PCL, the value of τ_2 is set to 0.05, and the number of sampling points S is determined to be 6. In U-AMLR, the size dimension of the cross-attention hidden layer for modal interactions is set to 512, and self-attention is set to be performed 4 times. For model training, we choose the Adam optimizer, initialize the learning rate to 1.0×10^{-5} , and use the strategy of cosine learning rate decay. The whole model is based on the PyTorch framework and trained on a GTX A6000 GPU with 80 epochs.

4.3 Comparison with state-of-the-art methods

On two publicly available datasets, we compare UANet with the baseline and other state-of-the-art methods.

As demonstrated in Table 1, on two publicly available datasets, our method achieves state-of-the-art performance on both the ICFG-PEDES and RSTPReid datasets. Specifically, on the RSTPReid dataset, each ID corresponds to five matching images and each image is associated with two textual descriptions, which results in a very large one-to-many correspondence in this dataset. Our approach employs an uncertainty-aware method for inter-modal alignment, which effectively models this complex association. As a result, compared with the state-of-the-art method IRRA [12], which is based on fixed-point embedding, we achieve the best performance currently available by improving the Rank-1, Rank-5 and Rank-10 metrics by 1.60%, 0.55% and 0.95%, respectively.

On the ICFG-PEDES dataset, since each person image is matched with only one text description, the challenge of one-to-many matching is not prominent. However, even so, our method still significantly improves the performance compared with that of the IRRA, with a gain 0.52% for the Rank-1 metric. These results prove that our method has a significant advantage in addressing one-to-one deterministic relationships as well.

Moreover, we measure the model's parameter size and inference time, and compare them with those of relevant open-source methods. Specifically, our approach is consistent with IRRA in using CLIP's visual and text encoders. However, to achieve fine-grained association reconstruction, we perform multiple image-to-text interactions, which increases the parameter size and inference time. However, in the testing phase, our approach only requires two feature encoders for feature extraction, without the need for additional auxiliary models and external knowledge, thus maintaining competitive inference time.

4.4 Ablation study

In this section, we perform a series of detailed ablation experiments on the RSTPReid dataset, aiming to validate the effectiveness of DCL, PCL, U-AMLR and regularized loss L_{reg} . Moreover, we compare the advantages of U-AMLR with other methods for constructing fine-grained associations. Importantly, we use similarity distribution matching loss [12] based on fixed-point embedding as our baseline.

Effectiveness validation for different settings. DCL and PCL achieve coarse-grained alignment between the two modalities from the perspectives of distributions and sampled points, respectively. Specifically, compared with fixed-point embedding-based similarity distribution matching loss, DCL leverages distribution modeling to yield a rich embedding space. As a result, DCL improves Rank-1 and Rank-5 by 1.70% and 1.75% over the baseline, indicating that distribution-based embeddings better capture cross-modal associations than fixed-point embeddings. PCL establishes cross-modal associations through inter-modal point contrastive learning, complementing DCL. Additionally, the intra-modal point contrastive learning further enhances distribution learning, yielding gains of 0.30%, 0.35%, and 1.30% on Rank-1, Rank-5, and Rank-10, respectively. The combined use of PCL and DCL leads to overall improvements of 2.05%, 2.10%, and 1.25% on Rank-1, Rank-5, and Rank-10 over the baseline. Furthermore, as shown in the rows 5 and 6 of the Table 1, using only PCL and other losses without DCL results in declines of 0.45%, 0.70%, and 0.20% on Rank-1, Rank-5, and Rank-10, respectively, further validating their synergy in establishing effective coarse-grained cross-modal associations.

Moreover, the L_{reg} is used to constrain the learning of the distribution, preventing it from degenerating into fixed-point embedding, which is crucial for ensuring uncertainty-aware learning. We validated the importance of this loss in our method, as shown in the rows 5 and 6 of Table 2. Without the L_{reg} , the distribution learning degenerates into fixed-point embedding learning, causing the model to fail to sample diverse embeddings and making it difficult to model complex cross-modal semantic associations. As a result, the performance on all the metrics significantly decreases, with Rank-1 and Rank-10 dropping by 1.15% and 0.85%, respectively. However, even so, the performance still significantly outperforms the baseline, owing to the introduction of other losses.

To more intuitively demonstrate the feature representations learned by the network after different levels of alignment, we visualize the attention between text and images, using heatmaps as illustrated in Fig. 3. After coarse-grained alignment (i.e., DCL+PCL), our method captures person-related regions, though its focus on finer details remains limited, with relatively low attention to areas, such as backpacks. However, following further fine-grained

Table 2 Ablation experiments with different settings on the RSTPReid dataset. Baseline refers to the method using fixed-point embeddings with similarity distribution matching loss, following Jiang and Ye. [12]. L_{reg} denotes the regularization loss for constraining distribution learning. DCL stands for distributed contrastive learning, PCL for point-wise contrastive learning, and U-AMLR represents the uncertainty-aware attribute-masked language reconstruction module

Method	Rank-1	Rank-5	Rank-10
Baseline	57.20	79.90	88.10
w/ L_{reg}	+DCL	58.95	81.65
	+DCL+PCL	59.25	82.00
	+PCL+U-AMLR	61.40	81.15
	+DCL+PCL+U-AMLR	61.85	81.85
w/o L_{reg}	+DCL+PCL+U-AMLR	60.70	81.65
		88.30	

alignment (i.e., U-AMLR), the model establishes more precise cross-modal associations, significantly enhancing attention to local details, with a noticeable increase in focus on the backpack. These results validate the effectiveness of our method.

Compared with PCL and DCL, U-AMLR further establishes the association between text entities and person parts through uncertainty-aware fine-grained alignment, resulting in a 2.60% improvement in the Rank-1 metric. However, since U-AMLR forces the reconstruction of masked words to match actual words, emphasizing the strong correspondence between text entities and body parts, this strong association may lead to overfitting on perfectly matched images. As a result, it somewhat weakens the diversity of matches, causing a slight decline in the Rank-5 and Rank-10 metrics. Nevertheless, compared with the baseline methods, the introduction of U-AMLR improves the overall performance of UANet, demonstrating its effectiveness.

Comparison of different ways of building fine-grained associations. As illustrated in Table 3, we design four experiments to construct fine-grained associations, including masked language modeling (MLM) and attribute masked language modeling (AMLR) using only mean alignment, as well as uncertainty-aware masked language model (U-MLM) and U-AMLR. The results indicate that both MLM and AMLR significantly enhance performance, confirming their effectiveness in building fine-grained associations between modalities. In particular, AMLR outperforms MLM by focusing on the masking of person attributes, thereby establishing its superiority with a more substantial performance enhancement. With the integration of uncertainty, U-MLM and U-AMLR exhibit 1.25% and 1.40% enhancements in Rank-1 metrics compared with MLM and AMLR, respectively, further validating the significant contribution of uncertainty-awareness in establishing precise fine-grained associations.

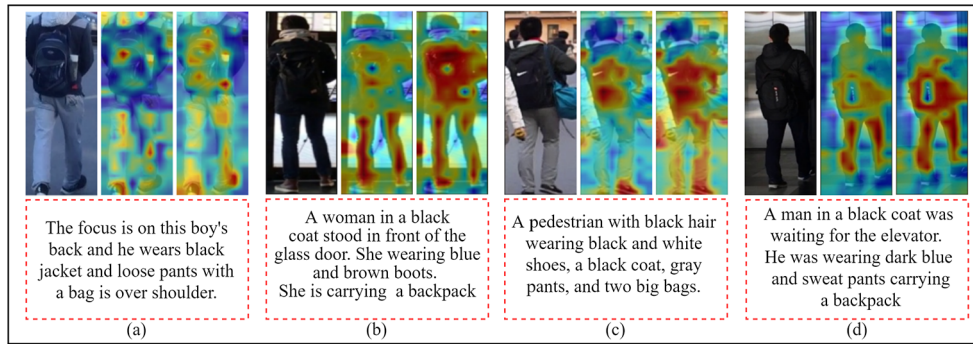


Figure 3 Comparison of cross-modal attention after different granularity alignments. The three columns for each person represent the original image, the attention map after coarse-grained alignment, and the attention map after further incorporating fine-grained alignment. The second row corresponds to the related text

Table 3 Comparison of different methods for constructing fine-grained associations. MLM denotes masked language model, AMLR denotes attribute masked language reconstruction and U-MLM denotes uncertainty-aware masked language model

Methods	Rank-1	Rank-5	Rank-10
Baseline+DCL+PCL	59.25	82.00	89.35
Baseline+DCL+PCL+MLM	60.15	80.80	89.00
Baseline+DCL+PCL+AMLR	60.45	82.10	87.90
Baseline+DCL+PCL+U-MLM	61.40	82.25	89.05
Baseline+DCL+PCL+U-AMLR	61.85	81.85	89.15

4.5 Parameter sensitivity analysis

To determine the number of sampling points S and the weights α to be used in the testing phase, we conduct a series of experiments in this section to select the optimal hyperparameters.

We first verify the effect of different weights α on the testing results on the RSTPreid and ICFG-PEDES datasets, and the results are shown in Table 4 and Table 5. When $\alpha = 0$, the matching probability is calculated entirely based on mean similarity, with Rank-1 accuracy of 60.10% and 62.34% achieved on the RSTPreid and ICFG-PEDES datasets, respectively. When $\alpha = 1$, the matching probability is determined only by the distribution similarity, with Rank-1 accuracy of 60.45% and 63.04% achieved on the RSTPreid and ICFG-PEDES datasets, respectively. Based on the experimental results on both datasets with different weights, when $\alpha = 0.7$, RSTPreid achieved the best Rank-1 and the second-best Rank-5, whereas ICFG-PEDES achieves the best Rank-1 and the second-best Rank-10. Therefore, we set $\alpha = 0.7$.

For selecting the number of sampling points, in consideration of memory constraints, we set the upper limit of it to 6. Accordingly, we conduct six sets of experiments on both datasets to evaluate the effect of different sampling point numbers on model performance, and the results are displayed in Table 6 and Table 7. The experimental results

show that the number of sampling points has a significant impact on model performance. When the number of sampling points S is set to 6, UANet achieves the best Rank-1 and Rank-10 scores on the RSTPreid dataset, and the best Rank-10 and second-best Rank-1 scores on the ICFG-PEDES dataset. Therefore, we set $S = 6$.

4.6 Qualitative results

Figure 4 demonstrates the comparison of the top 10 results between the UANet and IRRA models on the person image retrieval task. Although IRRA improves the retrieval performance by gradually aligning the fixed-point embedding from coarse to fine granularity, it has shortcomings in dealing with one-to-many relationship modeling, thus is unable to retrieve more persons that match descriptions. For example, for the retrieval results of the first two textual descriptions in Fig. 4, IRRA only retrieved one and three images matching the descriptions, respectively. However, our method retrieves five accurate results, which indicates that our method effectively solves the one-to-many matching problem and achieves a more comprehensive retrieval. Moreover, since MLM limits the ability of the model to build more accurate fine-grained association, it makes it difficult for IRRA to achieve accurate retrieval. For example, for the third textual description in Fig. 4, only the third out of the first ten retrieval results of IRRA are accurate. In contrast, our method achieves more accurate inter-modal fine-grained association than IRRA dose via U-AMLRL, and thus successfully matches the corresponding image at the first position. To better illustrate the advantages of U-AMLRL, we use orange bounding boxes in the retrieval results to highlight key regions corresponding to the orange descriptions in the text. For example, in the second text description, “white sneakers”, the top three retrieval results of IRRA fail to capture this detail, leading to incorrect matches. In contrast, our method successfully identifies and utilizes this attribute, achiev-

Table 4 Experimental results of different fusion weights α on the RSTPReid dataset

α	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Rank-1	60.10	60.40	60.40	61.05	61.50	61.55	61.85	61.85	<u>61.55</u>	61.00	60.45
Rank-5	82.05	81.75	81.65	81.55	81.80	81.75	81.85	<u>81.85</u>	81.75	81.75	81.90
Rank-10	88.85	89.05	88.80	88.80	88.90	89.10	89.15	89.15	89.40	89.40	<u>89.20</u>

Table 5 Experimental results of different fusion weights α on the ICFG-PEDES dataset

α	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Rank-1	62.34	62.65	62.67	63.06	63.37	63.56	<u>63.88</u>	63.97	63.56	63.28	63.04
Rank-5	80.12	80.25	<u>80.47</u>	80.56	80.16	80.06	80.18	80.27	80.22	80.10	80.05
Rank-10	84.55	85.07	85.16	85.28	85.12	85.24	85.38	85.54	85.47	<u>85.49</u>	85.29

Table 6 Performance comparison of the number of sampling points on the RSTPReid dataset

Number	1	2	3	4	5	6
Rank-1	60.55	60.45	<u>61.30</u>	60.65	60.6	61.85
Rank-5	81.05	83.35	81.35	81.45	81.85	<u>81.85</u>
Rank-10	88.35	<u>89.05</u>	87.45	88.20	88.95	89.15

Table 7 Performance comparison of the number of sampling points on the ICFG-PEDES dataset

Number	1	2	3	4	5	6
Rank-1	63.87	63.65	64.07	63.24	63.67	<u>63.97</u>
Rank-5	81.15	80.64	80.09	<u>80.76</u>	80.14	80.27
Rank-10	84.93	84.31	84.89	84.78	<u>85.08</u>	85.54



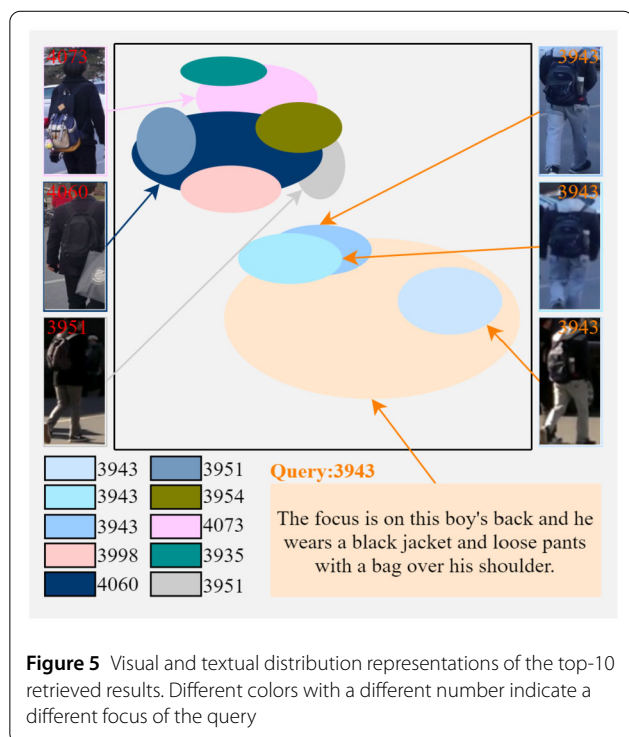
ing more accurate retrieval. Therefore, U-AMLR achieves richer and more accurate fine-grained alignment than that of the MLM dose.

5 Discussion

In the text-to-image person retrieval task, the uncertainty is often introduced by various factors, such as the polysemy of words, variations in image capture angles, and changes in lighting conditions. Additionally, the one-to-many cross-modal matching between text and images exacerbates this uncertainty. Traditional methods typically represent data as fixed-point embeddings and retrieve the visual features most similar to the query based on these points. However, these methods are only effective in deterministic contexts, as they consider only a single mapping of the cross-modal relationship. As a result, they fail to capture the complex cross-modal associations and struggle to address the issues arising from uncertainty.

To address this limitation, we propose UANet, which transforms fixed-point embeddings into distributional representations, enhancing the model's flexibility by sampling embeddings from this distribution. Additionally, through the proposed DCL, PCL, and U-AMLR, UANet establishes cross-modal multi-granularity associations. This approach expands the feature space, allowing for more robust cross-modal alignment. Moreover, when a textual description may correspond to multiple images of the same person, our method, compared with fixed-point embeddings, can capture various potential visual features under the same description—something fixed-point embeddings often fail to do. As a result, our method is better equipped to model complex cross-modal matching relationships.

As shown in Fig. 5, the distributional representation of each instance is presented, in which distributions with similar semantics are positioned closer to each other. For the three pedestrian images of identity 3943, one image differs significantly from the other two, which is reflected in their spatial distribution. Moreover, the distributions of the two similar images exhibit substantial overlap, indicating their close semantic relationship. This semantic variation is effectively modeled by our approach, as the text distribution can intersect with the image distributions of matching identities. Thus, our UANet method ensures a



rich understanding of instance semantics, enabling effective modeling of cross-modal matching associations.

Moreover, when selecting the hyperparameters, weight α and the number of sampling points S , we carefully consider their performance across different datasets and conduct systematic experiments to minimize the sensitivity of hyperparameter tuning to dataset variations. Specifically, we first adjust these two hyperparameters independently for each dataset and observe their impact on model performance. Experimental results show that when the weight parameter α is set to 0.7, the model achieves the best Rank-1 performance and second-best Rank-5 performance on the RSTPReid dataset, while on the ICFG-PEDES dataset, it achieves the best Rank-1 performance and second-best Rank-10 performance. This indicates that the choice of α demonstrates strong adaptability on both representative datasets. Regarding the choice of the number of sampling points S , the experimental results show that when S is set to 6, the model achieves the best Rank-1 performance and second-best Rank-10 performance on the RSTPReid dataset, and the best Rank-10 performance and second-best Rank-1 performance on the ICFG-PEDES dataset. This hyperparameter configuration exhibits good balance across different datasets, significantly enhancing the stability of the overall performance. These selected hyperparameters consistently deliver stable and near-optimal results across multiple datasets, greatly reducing the complexity of cross-dataset hyperparameter tuning. In the future, we plan to explore adaptive hyperparameter opti-

mization methods to dynamically adjust hyperparameters and better adapt to the characteristics of different datasets.

6 Conclusion

There are a many number of one-to-many correspondences in the text-to-image person retrieval task. Existing methods based on fixed-point embedding are often difficult to adequately capture this complex relationship, leading to erroneous retrieval results. To address this problem, we propose the uncertainty-aware coarse-to-fine alignment network, which significantly improves the retrieval performance by mapping fixed-point embedding to Gaussian probability distribution and performing coarse-to-fine-grained modality alignment from the perspective of distribution and sampling points. In particular, we propose the distribution contrastive learning and point contrastive learning tasks for inter-modal coarse-grained alignment, which not only enriches the embedding space, but also effectively models one-to-many correspondences. Moreover, for fine-grained inter-modal alignment, we design the uncertain-aware attribute mask language reconstruction task, which constructs accurate fine-grained associations by masking person attributes and reconstructing mask words in an uncertain-aware manner. Experiments on both public datasets validate the superiority of our method. However, a challenge to our approach is that as the number of sampling points increases, the model parameters also increase, which makes dense sampling difficult. To address this issue, future research will focus on improving the sampling strategy for more efficient sample points selection.

Abbreviations

CLIP, contrastive language-image pre-training; DCL, distribution contrastive learning; DGM, distribution generation module; GT, ground truth; MLM, masked language modeling; PCL, point contrastive learning; U-AMLR, uncertainty-aware attribute masked language reconstruction; UANet, uncertainty-aware coarse-to-fine alignment network.

Author contributions

YD wrote and revised the manuscript; ZC wrote and conducted the experiments; CL and JT provided guidance and supervision. All authors read and approved the final manuscript.

Funding information

This work was supported by the National Natural Science Foundation of China (No. 62376004), the Natural Science Foundation of Anhui Province (No. 2208085J18), and the University Synergy Innovation Program of Anhui Province (No. GXXT-2022-033).

Data availability

The datasets analyzed in this study, RSTPReid, are available in the repository at <https://github.com/NjtechCVLab/RSTPReid-Dataset>, and ICFG-PEDES can be accessed at <https://github.com/zifyloo/SSAN>.

Declarations

Competing interests

The authors declare no competing interests.

Author details

¹Information Materials and Intelligent Sensing Laboratory of Anhui Province, Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, School of Computer Science and Technology, Anhui University, Hefei, 230601, China.

²Information Materials and Intelligent Sensing Laboratory of Anhui Province, Anhui Provincial Key Laboratory of Security Artificial Intelligence, School of Artificial Intelligence, Anhui University, Hefei, 230601, China.

Received: 5 September 2024 Revised: 15 April 2025

Accepted: 16 April 2025 Published online: 29 April 2025

References

- Lei, J., Chen, X., Zhang, N., Wang, M., Bansal, M., Berg, T. L., & Yu, L. (2022). LoopITR: combining dual and cross encoder architectures for image-text retrieval. *arXiv preprint. arXiv:2203.05465*.
- Zhu, Y., Li, C., Tang, J., Luo, B., & Wang, L. (2022). RGBT tracking by trident fusion network. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(2), 579–592.
- Zhu, Y., Li, C., Luo, B., Tang, J., & Wang, X. (2019). Dense feature aggregation and pruning for RGBT tracking. In *Proceedings of the 27th ACM international conference on multimedia* (pp. 465–472). New York: ACM.
- García, J., Martinel, N., Micheloni, C., & Vicente, A. G. (2015). Person re-identification ranking optimisation by discriminant context information analysis. In *Proceedings of the IEEE international conference on computer vision* (pp. 1305–1313). Piscataway: IEEE.
- Guo, J., Yuan, Y., Huang, L., Zhang, C., Yao, J., & Han, K. (2019). Beyond human parts: dual part-aligned representations for person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 3641–3650). Piscataway: IEEE.
- Li, S., Xiao, T., Li, H., Wang, W., & Wang, X. (2017). Identity-aware textual-visual matching with latent co-attention. In *Proceedings of the IEEE international conference on computer vision* (pp. 1908–1917). Piscataway: IEEE Comput. Soc.
- Zhang, Y., & Lu, H. (2018). Deep cross-modal projection learning for image-text matching. In V. Ferrari, M. Hebert, C. Sminchisescu, & Y. Weiss (Eds.), *Proceedings of the 15th European conference on computer vision* (pp. 707–723). Cham: Springer.
- Wang, Z., Fang, Z., Wang, J., & Yang, Y. (2020). ViTAA: visual-textual attributes alignment in person search by natural language. In A. Vedaldi, H. Bischof, T. Brox, & J.-M. Frahm (Eds.), *Proceedings of the 16th European conference on computer vision* (pp. 402–420). Cham: Springer.
- Yan, S., Tang, H., Zhang, L., & Zhang, J. (2024). Image-specific information suppression and implicit local alignment for text-based person search. *IEEE Transactions on Neural Networks and Learning Systems*, 35(12), 17973–17986.
- Loper, E., & Bird, S. (2002). NLTK: the natural language toolkit. *arXiv preprint. arXiv:cs/0205028*.
- Jing, Y., Si, C., Wang, J., Wang, W., Wang, L., & Tan, T. Pose-guided multi-granularity attention network for text-based person search. In *Proceedings of the 34th AAAI conference on artificial intelligence* (pp. 11189–11196). Palo Alto: AAAI Press.
- Jiang, D., & Ye, M. (2023). Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2787–2797). Piscataway: IEEE.
- Ding, Z., Ding, C., Shao, Z., & Tao, D. (2021). Semantically self-aligned network for text-to-image part-aware person re-identification. *arXiv preprint. arXiv:2107.12666*.
- Zhu, A., Wang, Z., Li, Y., Wan, X., Jin, J., Wang, T., Hu, F., & Hua, G. (2021). DSSL: deep surroundings-person separation learning for text-based person retrieval. In H. T. Shen, Y. Zhuang, J. R. Smith, Y. Yang, P. Cesar, F. Metz, & B. Prabhakaran (Eds.), *Proceedings of the 29th ACM international conference on multimedia* (pp. 209–217). New York: ACM.
- Deng, Y., Li, C., Lu, A., Li, W., & Luo, B. (2022). Factory extraction from satellite images: benchmark and baseline. *Remote Sensing*, 14(22), 5657.
- Li, S., Xiao, T., Li, H., Zhou, B., Yue, D., & Wang, X. (2017). Person search with natural language description. In *Proceedings of the conference on computer vision and pattern recognition* (pp. 5187–5196). Piscataway: IEEE.
- Chen, T., Xu, C., & Luo, J. (2018). Improving text-based person search by spatial matching and adaptive threshold. In *Proceedings of the IEEE winter conference on applications of computer vision* (pp. 1879–1887). Piscataway: IEEE.
- Chen, Y., Huang, R., Chang, H., Tan, C., Xue, T., & Ma, B. (2021). Cross-modal knowledge adaptation for language-based person search. *IEEE Transactions on Image Processing*, 30, 4057–4069.
- Yan, S., Dong, N., Zhang, L., & Tang, J. (2023). CLIP-driven fine-grained text-image person re-identification. *IEEE Transactions on Image Processing*, 32, 6032–6046.
- Zheng, Z., Zheng, L., Garrett, M., Yang, Y., Xu, M., & Shen, Y. (2020). Dual-path convolutional image-text embeddings with instance loss. *ACM Transactions on Multimedia Computing Communications and Applications*, 16(2), 1–23.
- Layne, R., Hospedales, T. M., & Gong, S. (2014). Attributes-based re-identification. In *Person re-identification, advances in computer vision and pattern recognition* (pp. 93–117). Berlin: Springer.
- Vilnis, L., & McCallum, A. (2015). Word representations via Gaussian embedding. In Y. Bengio & Y. LeCun (Eds.), *Proceedings of the 3rd international conference on learning representations* (pp. 1–12). Retrieved March 10, 2025, from <http://arxiv.org/abs/1412.6623>.
- Chang, J., Lan, Z., Cheng, C., & Wei, Y. (2020). Data uncertainty learning in face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5709–5718). Piscataway: IEEE.
- Miech, A., Alayrac, J., Laptev, I., Sivic, J., & Zisserman, A. (2021). Thinking fast and slow: efficient text-to-visual retrieval with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9826–9836). Piscataway: IEEE.
- Chun, S., Oh, S. J., Rezende, R. S., Kalantidis, Y., & Larlus, D. (2021). Probabilistic embeddings for cross-modal retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8415–8424). Piscataway: IEEE.
- 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 *Proceedings of the 38th international conference on machine learning* (pp. 8748–8763). PMLR.
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2021). An image is worth 16x16 words: transformers for image recognition at scale. In *Proceedings of the 9th international conference on learning representations* (pp. 1–22). Retrieved March 10, 2025, from <https://openreview.net/forum?id=YicbFdNTTy>.
- Kingma, D. P., & Welling, M. (2014). Auto-encoding variational Bayes. In Y. Bengio & Y. LeCun (Eds.), *Proceedings of the 2nd international conference on learning representations* (pp. 1–14). Retrieved March 10, 2025, from <http://arxiv.org/abs/1312.6114>.
- Ji, Y., Wang, J., Gong, Y., Zhang, L., Zhu, Y., Wang, H., Zhang, J., Sakai, T., & 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* (pp. 23262–23271). Piscataway: IEEE.
- Wei, L., Zhang, S., Gao, W., & Tian, Q. (2018). Person transfer GAN to bridge domain gap for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 79–88). Piscataway: IEEE.
- Shu, X., Wen, W., Wu, H., Chen, K., Song, Y., Qiao, R., Ren, B., & Wang, X. (2022). See finer, see more: implicit modality alignment for text-based person retrieval. In S. Avidan, G. J. Brostow, M. Cissé, G. M. Farinella, & T. Hassner (Eds.), *Proceedings of the 17th European conference on computer vision* (pp. 624–641). Cham: Springer.
- Suo, W., Sun, M., Niu, K., Gao, Y., Wang, P., Zhang, Y., & Wu, Q. (2022). A simple and robust correlation filtering method for text-based person search. In S. Avidan, G. J. Brostow, M. Cissé, G. M. Farinella, & T. Hassner (Eds.), *Proceedings of the 17th European conference on computer vision* (pp. 726–742). Cham: Springer.
- Shao, Z., Zhang, X., Fang, M., Lin, Z., Wang, J., & Ding, C. (2022). Learning granularity-unified representations for text-to-image person re-identification. In *Proceedings of the 30th ACM international conference on multimedia* (pp. 5566–5574). New York: ACM.
- Ma, Y., Sun, X., Ji, J., Jiang, G., Zhuang, W., & Ji, R. (2023). Beat: bi-directional one-to-many embedding alignment for text-based person retrieval. In A. El-Saddik, T. Mei, R. Cucchiara, M. Bertini, D. P. Tobon Vallejo, P. K. Atrey, & M. S. Hossain (Eds.), *Proceedings of the 31st ACM international conference on multimedia* (pp. 4157–4168). New York: ACM.
- Yan, S., Dong, N., Liu, J., Zhang, L., & Tang, J. (2023). Learning comprehensive representations with richer self for text-to-image person re-identification. In A. El-Saddik, T. Mei, R. Cucchiara, M. Bertini, D. P. Tobon

- Vallejo, P. K., Atrey, & M. S. Hossain (Eds.), *Proceedings of the 31st ACM international conference on multimedia* (pp. 6202–6211). New York: ACM.
36. Shao, Z., Zhang, X., Ding, C., Wang, J., & Wang, J. (2023). Unified pre-training with pseudo texts for text-to-image person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 11140–11150). Piscataway: IEEE.
37. Liu, Y., Li, Y., Liu, Z., Yang, W., Wang, Y., & Liao, Q. (2024). Clip-based synergistic knowledge transfer for text-based person retrieval. In *Proceedings of the IEEE international conference on acoustics, speech, and signal processing* (pp. 7935–7939). Piscataway: IEEE.
38. Niu, K., Huang, L., Long, Y., Huang, Y., Wang, L., & Zhang, Y. (2024). Comprehensive attribute prediction learning for person search by language. *IEEE Transactions on Image Processing*, 33, 1990–2003.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Submit your manuscript to a SpringerOpen[®] journal and benefit from:

- Convenient online submission
- Rigorous peer review
- Open access: articles freely available online
- High visibility within the field
- Retaining the copyright to your article

Submit your next manuscript at ► [springeropen.com](https://www.springeropen.com)