Pattern Recognition 179 (2026) 113783

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Harnessing weak pair uncertainty for text-based person search

Jintao Sun [a], Zhedong Zheng [b,*], Gangyi Ding [a]

[a] School of Computer Science and Technology, Beijing Institute of Technology, China
[b] Faculty of Science and Technology, University of Macau, China

ARTICLE INFO

Keywords:
Text-based person search
Cross-modality
Uncertainty learning

ABSTRACT

In this paper, we study the text-based person search, which is to retrieve the person of interest via natural language description. Prevailing methods usually focus on the strict one-to-one correspondence pair matching between the visual and textual modality, such as contrastive learning. However, such a paradigm unintentionally disregards the weak positive image-text pairs, which are of the same person but the text descriptions are annotated from different views (cameras). To take full use of weak positives, we introduce an uncertainty-aware method to explicitly estimate image-text pair uncertainty, and incorporate the uncertainty into the optimization procedure in a smooth manner. Specifically, our method contains two modules: uncertainty estimation and uncertainty regularization. (1) Uncertainty estimation is to obtain the relative confidence on the given positive pairs; (2) Based on the predicted uncertainty, we propose the uncertainty regularization to adaptively adjust loss weight. Additionally, we introduce a group-wise image-text matching loss to further facilitate the representation space among the weak pairs. Compared with existing methods, the proposed method explicitly prevents the model from pushing away potentially weak positive candidates. Extensive experiments on three widely-used datasets, *i.e.*, CUHK-PEDES, RSTPReid and ICFG-PEDES, verify the mAP improvement of our method against existing competitive methods +3.06%, +3.55% and +6.94%, respectively. Code is available at https://github.com/JT-Sun/WPU-TBPS.

1. Introduction

Text-based person search is an extension of conventional image-based person re-identification (re-ID) [1,2], which is to retrieve the person of interest from a large pool of candidate images given text descriptions. In real-world scenarios, the image query of the target person usually is not accessible, while the test description is easy to obtain [3–5]. Therefore, more researchers resort to the text-based person search. The key underpinning this task is to mine the fine-grained information of images and texts, and establish the cross-modality alignment.

With the advancement of cross-modality learning, numerous deep learning approaches have been proposed, which can be broadly categorized into two directions. The first direction focuses on data augmentation, primarily by generating additional data and employing the pretrain-finetune paradigm. For example, based on a pre-trained model, Jiang and Ye [6] propose a momentum distillation cross-modal method using four datasets for pre-training, which enables the model to utilize larger noisy datasets, thereby improving learning under noisy supervision. Differently, Yang et al. [3] propose a large-scale image-text dataset with high similarity to the target dataset, constructed using a diffusion model, which addresses the challenges of image collection and time-consuming text annotation. The second direction emphasizes metric learning, aiming to design more effective loss functions to better exploit multimodal information within the data, such as contrastive loss and cross-modal matching loss [7]. For instance, Shao et al. [8] propose to use cross entropy loss and ranking loss to get the total loss of multimodal shared storage dictionaries. Bai et al. [9] incorporate relation-aware loss and sensation-aware loss, enabling the model to focus more on the details of image-text pairs and learn cross-modal features with finer granularity.

We observe that there is an inherent limitation in text-based person retrieval metric learning, which primarily relies on strict one-to-one contrastive learning on positive pairs. The weak positive pairs are usually disregarded due to the annotation discrepancy. The annotator only observes one view (camera angle) of the target person, and cannot provide a comprehensive description. As illustrated in Fig. 1, images of the same person usually be annotated with variations in local details, resulting in a mismatch between the image and text descriptions across different perspectives. Consequently, strict one-to-one matching discards these weak positive pairs during training, overlooking potentially valuable information. However, we argue that weak positive pairs play a crucial role in enhancing the model ability to capture shared

* Corresponding author.
E-mail addresses: 3120215524@bit.edu.cn (J. Sun), zhedongzheng@um.edu.mo (Z. Zheng), dgy@bit.edu.cn (G. Ding).

https://doi.org/10.1016/j.patcog.2026.113783
Received 12 November 2025; Received in revised form 14 April 2026; Accepted 15 April 2026
Available online 20 April 2026

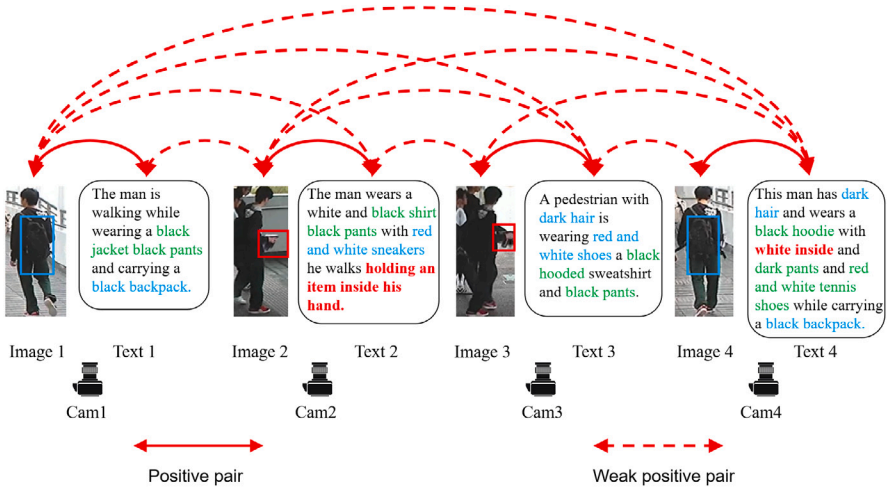


Fig. 1. Here we show the discrepancy between weak positive pairs (dotted arrows) and positive pairs (solid arrows) of the same identity. For instance, there are four image-text pairs of one person. **Green** shows a shared description among four texts, **blue** denotes the description discrepancy, and **red** indicates the unique description. We can observe that the text description is strongly related to camera views instead of only depending on the identity. Such difference is mainly due to the text annotation process, where annotators can only see a single view of the person. Considering such matching difficulty, most previous methods, thus, take advantage of the positive pairs for strict one-to-one matching (solid arrows), while disregarding the weak positive pairs (dashed arrows). In contrast, we mine the relation among weak positive image-text pairs and explicitly harness such weak positive pairs to learn discriminative cross-modality matching. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

patterns across different views of the same individual. These shared patterns should be positioned closer to the anchor points in the feature space than negative pairs, enabling the model to better generalize across varying perspectives. Despite this, the potential of weak positive relationships remains under-explored in current methodologies.

To take full usage of weak positive pairs from the dataset, we propose a new uncertainty-aware learning method to solve some unutilized data problems. There are two steps in total. First, we adjust uncertainty for the comparison learning of data features that can be most affected by data underutilized. Under the condition of keeping a one-to-one correspondence comparison of positive data features, we leverage the auxiliary information of weak positive pairs. The feature similarity of images and texts of weakly positive pairs are calculated respectively, and parameters are set to harness and adjust the feature contrast of weakly positive pairs, which is added to the total loss calculation, thus improving the model ability to learn from data. In the second step, we exploit the impact of data representation space in the construction of negative pairs during metric learning to both increase the quantity and difficulty of negative pairs. Additionally, our group-wise matching method makes full use of the information of weak positive pairs to make the representation space distribution of the model more reasonable. This approach enables the calculation of the matching loss for 1 pair of positive samples, 2 pairs of weak positive samples, and 6 pairs of negative samples. Finally, our uncertainty-aware approach greatly improves the learning accuracy of the model without adding additional modules. In summary, our contributions are:

- We observe a limitation in the existing text-based person search training, which stems from the strict one-to-one correspondence contrastive learning approach. To address this issue, we propose an uncertainty learning-based method that effectively leverages underutilized weak positive pairs. Specifically, our method incorporates uncertainty into the feature comparison process, enabling the model to leverage weak positive pairs rather than discarding them. As a minor contribution, we also propose a Group-wise Image-Text Matching (GITM) loss, which facilitates the matching of weak positive pairs in a group-wise manner.
- Extensive experiments verify that our uncertainty-aware method, considering weak positive pairs during training, recalls more positive candidates to the top ranking. In particular, our approach outperforms competitive methods, *e.g.*, RaSa and APTM, by 3.06%, 3.80%, 6.94% mAP and 5.53%, 3.55%, 7.01% mAP on CUHK-PEDES, RSTPReid, and ICFG-PEDES, respectively.

2. Related work

2.1. Text-image person search

Text-based person search constitutes a challenging fine-grained cross-modal retrieval task, prompting the emergence of diverse methodologies in recent years. Existing approaches are primarily categorized into two paradigms: those leveraging explicit cross-modal attention interactions to enhance regional-word/phrase correspondence and predict image-text matching scores through sophisticated attention mechanisms [10–12], which improve inter-modal fusion at the cost of elevated computational complexity, and lightweight alternatives that forgo such interactions by learning aligned representations within a shared feature space via carefully designed architectures and objectives [13,14]. Early efforts, such as dual-path convolutional frameworks [14], exploit end-to-end supervision to derive modality-specific features, while more recent advances incorporate vision-language pretraining to yield robust representations [2,3,9], often augmented by attribute prompt learning, relation-aware modeling, or multi-attribute datasets like MALS to facilitate fine-grained alignment. Nevertheless, prevailing methods largely overlook the rich auxiliary supervisory signals embedded in weak positive pairs within the datasets. In contrast, the proposed approach introduces uncertainty learning and regularization to refine both contrastive learning and image-text matching objectives, thereby fully harnessing weak positive information to enhance the discriminative capacity of positive pair representations.

2.2. Uncertainty learning

Uncertainty quantification has become increasingly important in data-driven methods as datasets grow larger and demands for model reliability intensify. Kendall and Gal [15] provide a foundational taxonomy that decomposes predictive uncertainty into aleatoric uncertainty, which captures irreducible data-inherent noise, and epistemic uncertainty, which reflects model parameter ambiguity arising from

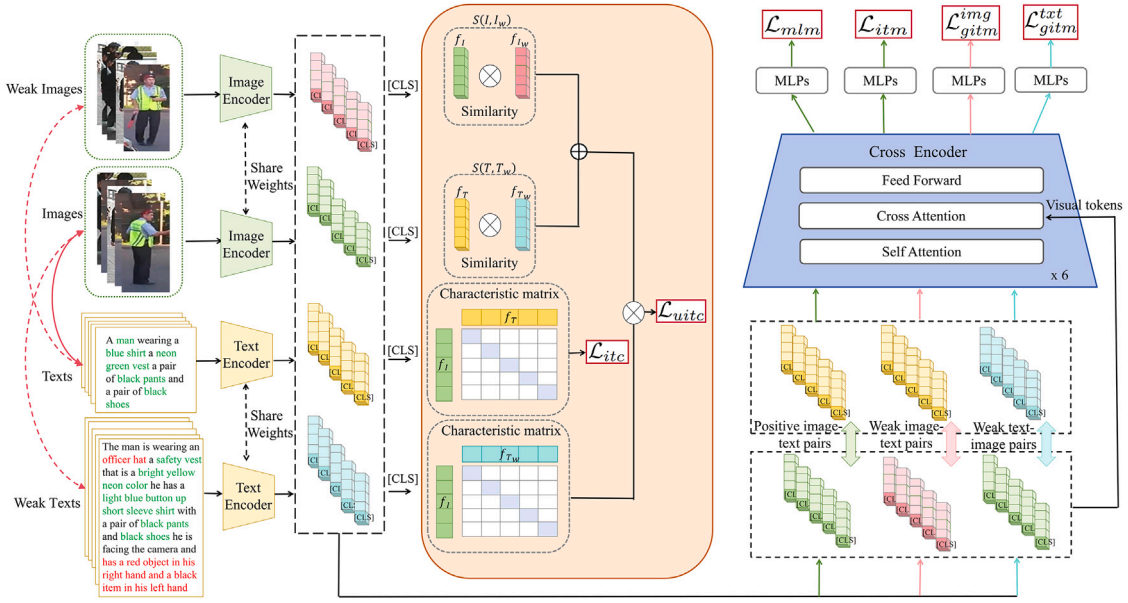


Fig. 2. An architecture overview of our approach. Firstly, weakly positive text and weakly positive image corresponding to the anchor image and text are randomly selected from the dataset according to the same ID. Then we send all sampled data to the image/text encoder to obtain the corresponding features. Secondly, we compute the ITC loss between images and texts. For the anchor image-text pairs, we calculate the original contrastive learning loss, *i.e.*, $\mathcal{L}_{itc}(I,T)$ via the aggregated embeddings of [CLS]. For the weak image-text pairs, we derive the contrastive objective, *i.e.*, $\mathcal{L}_{itc}(I,T_w)$, and further regularize this term with uncertainty estimation as $\mathcal{L}_{uitc}$. The proposed uncertainty-aware contrastive loss $\mathcal{L}_{uitc}$ adaptively adjust the metric learning between weakly positive texts and images. Thirdly, ITM loss and MLM loss are calculated based on the features of image-text pairs. For weak positive counterparts, we introduce Group-wise Image-Text Matching (GITM) loss to facilitate the representation learning.

limited or insufficient training data and can be reduced through additional observations or targeted refinements. Aleatoric uncertainty has been extensively explored in computer vision tasks, including image retrieval [16,17], classification [18], and segmentation [19], with approaches such as explicit noise injection into features [20,21], Monte Carlo estimation of distributional similarity [22], and loss-variance-aware reweighting [16]. In contrast, epistemic uncertainty modeling commonly leverages Bayesian frameworks [23,24], with Monte Carlo Dropout [24] serving as a practical approximation by introducing stochasticity during inference; recent works further integrate cross-modal biases [6] and dynamic uncertainty-guided learning. Multi-granularity annotation strategies combined with Gaussian noise simulation have also been employed to explicitly address aleatoric effects in composed image retrieval [17]. Building on the insights from previous work, this study fully leverages the information contained in weak positive pairs within the dataset. There are two fundamental differences between previous work and our approach: (1) We do not introduce additional modules or parameters to simulate uncertainty using noise. Instead, we harness the weak positive pairs, which describe different image-text pairs for the same ID, as a source of uncertainty to support the model contrastive learning of image-text features. (2) We employ uncertainty regularization, and apply a group-wise strategy to incorporate the semantic information from weak positive pairs, thereby enhancing the image-text matching process.

3. Methodology

3.1. Preliminaries

We employ the prevailing APTM framework [3] as our baseline model. This framework comprises two primary phases: pre-training on a synthesized dataset and fine-tuning on downstream datasets. During the pre-training phase, Attribute Prompt Learning (APL) and Text Matching Learning (TML) are employed to capture shared knowledge relevant to text-based person search and pedestrian attribute recognition. In the fine-tuning phase, downstream datasets are utilized to further optimize the model parameters. The baseline comprises three encoders: image encoder, text encoder, and cross encoder, along with two MLPs-based headers. In this work, we do not change the pre-training phase and only apply our uncertainty estimation and uncertainty regularization methods in the finetune phase. We do not pursue the network contribution in this work. Therefore, we adopt the common backbone for a fair comparison. The influence of the original image encoder and the replaced image encoder on the results are compared in detail in the ablation studies. The [CLS] embedding represents the entirety of the image/text. The cross-encoder integrates image and text representations for prediction tasks, thereby discerning their semantic relationship.

3.2. Network structure

In this paper, we do not pursue the deployment of complex network structures but instead offer a new learning strategy. We mainly follow the existing work [3,4] to construct the network for a fair comparison. Given that datasets contain images and text descriptions of the same person from different perspectives, previous work often overlooks the weak correlation between images and text under the same ID. However, the auxiliary information of weak positive pairs can significantly enhance the model ability to learn a more comprehensive representation of the description of person. Therefore, we mainly consider the uncertainty of image-text pairs describing the same person in the dataset. As shown in Fig. 1, the dataset usually contains multiple image-text pairs describing the same ID. The previous method relies solely on one-to-one image-text pairs for training, overlooking the relationship between the current image and the texts captured from another perspective, as well as the connection between the current text and images taken from different angles. Consider that these are not exactly the same texts and images. If the text is biased relative to the source image, or if the target image contains more perspective information than the source text, the system will supplement this uncertain information. In the training stage, in order to reduce the visual information bias and text description bias of the same person from different perspectives, in this work,

we mainly study the uncertainty in cross-modal data matching. We show a brief pipeline in Fig. 2. A one-to-one correspondence positively correlated image-text pair (I,T) and weakly correlated image I_w and text T_w randomly selected for this image-text pair (I,T) are extracted from the dataset. Our model extracts the features of image I to obtain f_I and f_{I_w}. Text encoder extracts text features T to get f_T and f_{T_w}, and carries out contrastive learning for image-text features. Meanwhile, weak positive pairs uncertainty are added to the contrastive learning. The obtained image embeds and text embeds, as well as hard negative pairs obtained by contrast learning based on image text features and hard negatives, increased based on our group-wise method, are sent to cross encoder together. The specific cross-modal data uncertainty approach is detailed in the following two sections.

3.3. Uncertainty estimation

Motivations. Existing text-based person search datasets provide image and text descriptions from different perspectives for the same individual. However, current methods often focus solely on one-to-one image-text pairs, which frequently fail to fully capture the characteristics of a person due to feature deviations across different perspectives. Learning features from strictly one-to-one image-text pairs can be limiting, as the method only leverage one-to-one image-text pairs usually overlook certain characteristics, leading the model to perform retrieval based on incomplete information.

First, we define the one-to-one corresponding image-text pairs (I,T) of the input model in the dataset as positive pairs, the unmatched image-text pairs as negative pairs, and the weakly positive correlation image-text pairs (I_w,T_w). I_w is the weakly positive image relative to T obtained by random extraction according to the current one-to-one corresponding image-text pair (I,T) in the case of describing the same person (same ID), and T_w is the weakly positive text description relative to I obtained by random extraction under the case of the same ID. The existing methods usually adopt Image-Text Contrastive Learning (ITC) to distinguish positive and negative pairs. Given a matched pair (I,T) we initially extract their respective representations f_I and f_T. We denote the set of all matched image-text pairs in a mini-batch as B. The matching score can be simply formulated as:

$$S(I,T)=\frac{\exp\left(\cos\left(f_I,f_T\right)/\tau\right)}{\sum_{i=1}^{B}\left(\exp\left(\cos\left(f_I,f_{T^i}\right)/\tau\right)\right)}, \tag{1}$$

where τ is a learnable temperature parameter, $\cos(\cdot,\cdot)$ means the cosine similarity, and exp denotes the exponential function. Similarly, given the text and a batch of images, we calculate the matching score of the paired image $S(T,I)$. The ability to differentiate learning is added to the final loss calculation. The ITC loss is defined as follows:

$$\mathcal{L}_{itc}(I,T)=-\mathbb{E}\left(\log S(I,T)+\log S(T,I)\right). \tag{2}$$

We compute cosine similarity on L2-normalized embeddings (as in standard retrieval baselines), hence $\cos(\cdot,\cdot)\in[-1,1]$. Moreover, the ITC objective in Eq. (1) already includes a temperature parameter τ in the logits $(\cos(\cdot,\cdot)/\tau)$, which controls the sharpness of the softmax distribution and stabilizes gradients; τ is learnable in our implementation, consistent with the baseline.

In order to take full advantage of the annotation, we propose an uncertainty estimation method. We integrate the weak positive pair into the text image alignment. Since our task is text-based image retrieval, the first step of our uncertainty estimation method is to extract the features f_I of the positive image I and the text feature f_{T_w} of the weak text T_w is with the same ID but annotated based on a different viewpoint. Then we calculate the obtained image feature f_I and weak text feature f_{T_w} matching score, refer to Eq. (1). Therefore, we can obtain the matching score $S\left(I,T_w\right)$ of positive image features and weak positive text features with the same ID. Similarly, we can get the matching score $S\left(T,I_w\right)$ of positive text features and weak positive image features with the same ID. According to the conventional ITC method Eq. (2), $\mathcal{L}_{itc}\left(I,T_w\right)$ is calculated for subsequent uncertainty regularization. Next, we extract the corresponding image and text features f_{I_w} and f_{T_w} according to the input weak positive pair $\left(I_w,T_w\right)$. The image feature f_I and the text feature f_T of the positive image-text pair (I,T) with the same ID. Then, we calculate the similarity between f_I and f_{I_w}, f_T and f_{T_w} respectively. We define $\cos\left(f_I,f_{I_w}\right)$ as the similarity calculated between f_I and f_{I_w}. Similarly, $\cos\left(f_T,f_{T_w}\right)$ is derived in the same manner. We define the consistency score s_w as the sum of intra-modality similarities between the anchor and weak positive samples. To measure the semantic discrepancy across views, we compute the uncertainty u_w via an exponential transform:

$$s_w=\frac{1}{2}\Big(\cos(f_I,f_{I_w})+\cos(f_T,f_{T_w})\Big),\qquad u_w=exp\left(-s_w\right). \tag{3}$$

Here, f_I and f_T denote the image/text features of an anchor sample, while f_{I_w} and f_{T_w} are the corresponding features of its paired sample w (e.g., another view or a weakly matched counterpart). $\cos(\cdot,\cdot)$ is the cosine similarity. Since cosine similarity is bounded in $[-1,1]$, we have $s_w\in[-1,1]$ and thus $u_w=\exp(-s_w)\in[e^{-1},e^{1}]$. Therefore, u_w is strictly positive and bounded, and larger u_w indicates lower cross-view consistency (higher ambiguity) in weak-pair matching. Our uncertainty estimation method fully leverages auxiliary information of images and texts and utilizes annotation information in data sets to supplement the information between images and texts, so as to enhance the model ability to accurately capture and interpret the nuanced relationships between visual and textual data, leading to more precise and reliable feature extraction and matching.

Discussion. What is the advantage of uncertainty estimation in feature learning? Uncertainty estimation is pivotal in enhancing the robustness of feature learning by simulating scenarios in which different views or descriptions of the same person are matched with alternative views or textual descriptions under the same ID. This approach effectively mitigates discrepancies that commonly arise between varying descriptions of the same individual, thereby reducing deviations that can negatively impact model performance. Moreover, uncertainty estimation addresses the challenge of overfitting associated with strict one-to-one matching, enabling the model to generalize more effectively and to reason in a manner more aligned with human cognition. By diminishing the homogeneity and reducing the deviations between image and text descriptions, uncertainty estimation enhances the model ability to learn more comprehensive and accurate representations. This improvement not only bolsters the model robustness against variations in descriptions but also leads to superior performance in real-world scenarios, where perfect alignment between descriptions and images is often lacking. Our experimental results verify that the proposed method significantly outperforms existing approaches, particularly in terms of mean Average Precision (mAP), highlighting its effectiveness in addressing the complexities inherent in text-based person search tasks.

Applicability and limitations. The proposed uncertainty learning is most effective when weak positives are informative but imperfectly aligned, as in TBPS datasets with multi-view annotations. It may yield limited gains when weak positives become noisy supervision (e.g., severe occlusion/viewpoint changes, non-overlapping attribute descriptions, or annotation mismatch), where a weak positive can behave close to a pseudo-negative in the contrastive space. In practice, Fig. 6 provides a diagnostic: high-u queries/pairs exhibit higher retrieval risk, and a large mass of high-u weak pairs indicates that weak-positive supervision is less reliable.

3.4. Uncertainty regularization

Motivations. Since conventional ITC and ITM loss only rely on one-to-one correspondence positive pairs in the data set, and do not maximize the utilization of the weak positive information in the labeling, in order to make full utilize of this information to assist us

in training the model, we propose uncertainty regularization to adjust the model learning of ITC loss, and group-wise metric learning for ITM loss so that the model can better grasp the information in the data. The existing work aims at negative pair mining in ITM loss. The conventional ITM method only compares and learns (I, T) based on one-to-one correspondence mapping of images and texts in the input dataset to obtain negative pair construction. Such construction will ignore some useful auxiliary weak positive instances and have less structural information. The difficulty of the negative pair is only determined by the one-to-one correspondence image-text pair and the batch size, and much auxiliary information is not considered in the weak positive pair, so the difficulty of the negative pair has certain limitations. Moreover, in existing works, the ratio of positive pair to negative pairs is fixed at 1:2, where for each input image-text pair (I, T), the image is used to find a negative text, and the text is used to find a negative image. Consequently, the number of negative pairs is fixed, which limits the quantity and diversity of negative pairs available.

To solve the above problem, in particular, we exploit to the fullest extent the image features and text features of weak positive pairs in the task and obtain the loss of images and weak positive texts through uncertainty estimation in Section 3.3. The similarity between the extracted image and text features of positive pairs and those of weak positive pairs with the same ID, denoted as $\cos\left(f_I, f_{I_w}\right)$ and $\cos\left(f_T, f_{T_w}\right)$, respectively, we can obtain an uncertainty-aware ITC loss:

$$\mathcal{L}_{uitc} = \frac{\mathcal{L}_{itc}\left(I, T_w\right)}{\gamma \times u_w} + \gamma \times u_w, \tag{4}$$

where γ is a learnable parameter used to adjust the uncertainty adjustment, and $\mathcal{L}_{itc}\left(I, T_w\right)$ is calculated by inputting image feature f_I and with the text feature f_{T_w} of the weak positive of this image into the original ITC loss calculation and applying Eq. (2) to calculate. The uncertainty u_w is derived via an exponential mapping of the average intra-modality similarities between the weak positive samples and the anchor pairs, as formulated in Eq. (3). In implementation, we stop the gradient through u_w when applying uncertainty regularization, so that u_w serves only as a detached reliability weight and the network cannot reduce Eq. (4) by directly manipulating the weighting path.

We note that ITC loss only focuses on the one-to-one matching of image text pairs, and as we mentioned earlier, these one-to-one correspondence text pairs are not comprehensive when describing a person. Therefore, in this work, inspired by uncertainty, we propose an uncertain regularization to optimize the existing ITC loss. In fact, the existing ITC loss is a special case of our uncertainty adjustment, when the input only has a one-to-one correspondence text pair corresponding to one person and no other multiple descriptions or perspectives.

Image-text Matching (ITM) Learning is a binary classification method for predicting whether an input image and text match. Eq. (1) in ITC is used to calculate the similarity of input images and text features and select the unpaired image with the highest similarity to each text as the hard negative. Similarly, we could select the unpaired text with the highest similarity as the hard negative for images. Such 1 pair of positive samples and 2 pairs of negative samples go through the cross-encoder with ITM loss:

$$\mathcal{L}_{itm} = \mathbb{E}\left[p \log \hat{p}(I, T) + (1-p) \log\left(1 - \hat{p}(I, T)\right)\right], \tag{5}$$

where p is 1 if (I, T) is matched, 0 otherwise, and $\hat{p}$ is the estimated match score of image-text pairs calculated by an MLPs with Sigmoid activation.

For each anchor matched pair (I_i, T_i) in a mini-batch, we form a group by additionally sampling one weak-positive image I_i^w and one weak-positive text T_i^w from the same identity, yielding two weak-positive pairs (I_i, T_i^w) and (I_i^w, T_i) (together with the original strong pair (I_i, T_i)). Hard negatives are mined within the current mini-batch based on cosine similarity of the current embeddings $s_{ij} = \cos(f_i^I, f_j^T)$: for each anchor identity, we only consider unpaired samples from different identities ($\mathrm{id}(j) \neq \mathrm{id}(i)$) and select the top-$K$ most similar ones as hard negatives in both image→text and text→image directions. In our implementation, all constructed pairs of a mini-batch are flattened and evaluated by the same ITM classifier, and the loss is computed by averaging over the constructed pairs per anchor group.

To make the above construction explicit, we denote the binary ITM log-likelihood term as $\ell(I, T, p) = p \log \hat{p}(I, T) + (1-p)\log(1 - \hat{p}(I, T))$, where p=1 for matched pairs and p=0 for negatives. Let $\mathcal{N}_i^T$ be the indices of top-K hard negative texts for I_i, and $\mathcal{N}_i^I$ be the indices of top-K hard negative images for T_i mined in the batch. Then the two GITM auxiliary branches in Eqs. (8)–(9) are instantiated by averaging one weak-positive term and its mined negatives as follows:

$$\mathcal{L}_{gitm}^{txt} = \mathbb{E}_i\left[\frac{1}{1+K}\left(\ell(I_i, T_i^w, 1) + \sum_{j \in \mathcal{N}_i^T} \ell(I_i, T_j, 0)\right)\right], \tag{6}$$

$$\mathcal{L}_{gitm}^{img} = \mathbb{E}_i\left[\frac{1}{1+K}\left(\ell(I_i^w, T_i, 1) + \sum_{j \in \mathcal{N}_i^I} \ell(I_j, T_i, 0)\right)\right]. \tag{7}$$

With the standard ITM term on the strong pair (I_i, T_i) providing two directional mined negatives, and the two weak-positive branches each attaching K mined negatives, the overall per-anchor ratio becomes three matched pairs versus $2+2K$ mined negatives (neg3v4 uses K=1, while neg3v6 uses K=2).

Implementation summary of GITM group. For re-implementation clarity, Table 1 summarizes the per-anchor construction used by GITM. It does not introduce any new component; it only restates the weak-positive sampling, hard-negative mining, and loss evaluation procedure in a compact implementation-oriented form.

To further facilitate the learning from weak positive pairs, we propose Group-wise Image-Text Matching (GITM) loss, shown in Fig. 3. Different from the original ITM loss, we include more negative pairs as well as the weak positive pairs. To sample the pair, we calculate the similarity between the positive image and the weak positive text, and vice versa. We select multiple hard negative samples and weak positive samples. Similarly, the image-text GITM loss $\mathcal{L}_{gitm}^{txt}$ and text-image GITM $\mathcal{L}_{gitm}^{img}$ based on the weak positive pairs can be formulated as:

$$\mathcal{L}_{gitm}^{txt} = \mathbb{E}\left[p \log \hat{p}(I, T_w) + (1-p) \log\left(1 - \hat{p}(I, T_w)\right)\right], \tag{8}$$

$$\mathcal{L}_{gitm}^{img} = \mathbb{E}\left[p \log \hat{p}(I_w, T) + (1-p) \log\left(1 - \hat{p}(I_w, T)\right)\right], \tag{9}$$

where p is 1 if (I, T_w) or (I_w, T) is matched, 0 otherwise, and $\hat{p}$ is the match score prediction of image and weak text pairs or weak image and text pairs calculated by an MLPs with Sigmoid activation. In summary, the final loss can be formulated as:

$$\mathcal{L}_{total} = \mathcal{L}_{mlm} + \mathcal{L}_{itc} + \mathcal{L}_{itm} + \alpha\mathcal{L}_{uitc} + \beta(\mathcal{L}_{gitm}^{txt} + \mathcal{L}_{gitm}^{img}), \tag{10}$$

where α controls the contribution of the uncertainty-aware contrastive term $\mathcal{L}_{uitc}$ in the multi-loss objective, and β controls the strength of the auxiliary GITM regularization terms $\mathcal{L}gitm^{txt}$ and $\mathcal{L}gitm^{img}$ in the multi-loss objective. We select α via the dedicated sweep in Table 10 under the same setting, and use the best-performing value ($\alpha = 0.5$) for all remaining experiments. For the GITM branch, we keep a small fixed coefficient ($\beta = 0.1$) so that $\mathcal{L}_{gitm}^{txt}$ and $\mathcal{L}_{gitm}^{img}$ act as auxiliary stabilizers rather than dominating the optimization when combined with $\mathcal{L}_{itc}$ and $\mathcal{L}_{itm}$. Importantly, the systematically validated factor for GITM is the *group construction and negative count* (*i.e.*, the positive-to-negative ratio induced by hard-negative mining), which is ablated in Table 7 by comparing neg3v4 and neg3v6. This confirms that increasing the number of hard negatives per group is the primary driver for the additional gains brought by GITM, while the loss-weight is kept fixed across settings.

Why is u_w used in ITC but not in GITM? The two branches serve different purposes in our framework. The uncertainty score u_w

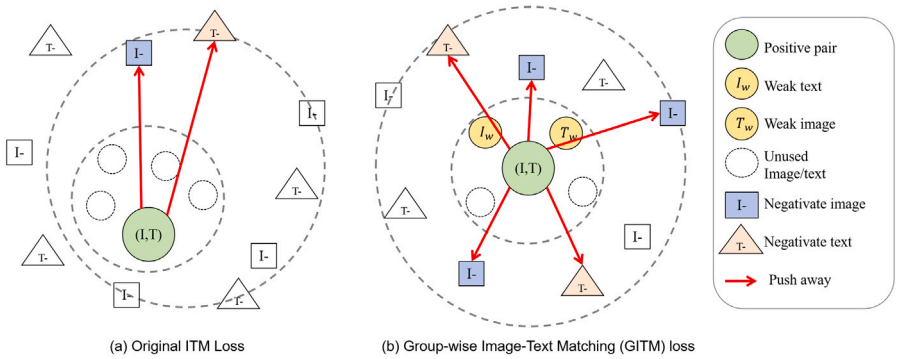


Fig. 3. Intuitive illustration of our Group-wise Image-Text Matching (GITM) loss approach. (a) In the conventional ITM loss calculation, one pair of positive pairs and two negative pairs are used, resulting in limited negative pair diversity. This lack of diversity leads to a skewed representation space distribution, potentially reducing the accuracy of the model performance. Additionally, conventional ITM does not fully leverage all available image-text data, causing semantic deviations between images and texts captured from different perspectives. These deviations can further hinder the ability of the model to effectively learn positive and negative pair matching. (b) Our Group-wise Image-Text Matching (GITM) approach introduces weak positive pairs, allowing the model to learn a more robust latent space for positive pairs while accounting for more diverse scenarios. By utilizing a larger and more diverse set of negative pairs, GITM increases both the number and difficulty of these pairs, resulting in a more evenly distributed representation space and, consequently, enhanced learning accuracy.

Table 1
Compact implementation summary of GITM for one anchor pair (I_i, T_i).

Step	Summary
Anchor pair	Start from one strong matched image-text pair (I_i, T_i).
Weak positives	Sample one weak-positive image I_i^w and one weak-positive text T_i^w from the same identity, forming two weak-positive pairs (I_i, T_i^w) and (I_i^w, T_i).
Hard negatives	Within the current mini-batch, only consider unpaired samples from different identities. The standard ITM term on the strong pair (I_i, T_i) provides two directional hard negatives (one image→text and one text→image). In addition, for the two weak-positive branches, we mine the top-K most similar negatives in each direction based on the current embeddings.
Per-anchor group	The final group contains three matched pairs, i.e., (I_i, T_i), (I_i, T_i^w), and (I_i^w, T_i), together with $2+2K$ mined negatives in total. Thus, neg3v4 uses $K=1$, while neg3v6 uses $K=2$.
Loss evaluation	Flatten all constructed pairs in the mini-batch, evaluate them using the same ITM classifier, and average the loss over anchor groups.

is introduced in the ITC branch to regulate the *pair-level ambiguity* of weak positives in the contrastive space: when a weak pair is less reliable, its contribution is softly reduced, whereas more consistent weak pairs provide stronger auxiliary supervision. By contrast, GITM is designed as an auxiliary *binary matching* branch to enrich the *group structure* by explicitly introducing weak positives together with more and harder negatives. Therefore, in GITM, weak positives are assigned the standard matched label, while the key design factor is the group construction/negative count rather than uncertainty reweighting. In this way, uncertainty-aware soft regulation is handled in ITC, whereas group-wise structural enrichment is handled in GITM.

Discussion. (1) Why is the proposed ITC loss based on uncertainty adjustment effective? The proposed method introduces additional uncertainty through weak orthogonal modeling. If the uncertainty value is high, it denotes the semantic gap between text and image in the weak positive pairs is large. Therefore, we automatically decrease the loss to mitigate the negative impact. If the uncertainty is small, we leverage the weak positive pairs as the positive pairs as auxiliary supervision. **(2) What is the motivation for GITM loss?** Group-wise Image-Text Matching (GITM) enables us to fully leverage the image and text features of weak positive pairs (as shown in Fig. 3). This strategy allows the model to extract a more comprehensive relationship between multiple pairs. In experiment, we increase the number of positive and negative pairs, altering the ratio from 1:2 to 3:6 (comprising one pair of strong positive samples and two pairs of weak positive samples, and six negative pairs). We observe that larger group-wise metric learning boosts the diversity of negative pairs and thus, enhances the model ability to discriminate between more negatives.

4. Experiment

4.1. Datasets and evaluation protocol

Datasets. We employ the synthetic dataset MALS [3] for pre-training, which comprises 1,510,330 image-text pairs, each annotated with relevant attribute labels. We validate our method on three benchmark datasets. For fine-tuning and evaluation, we utilize widely-used datasets: CUHK-PEDES [25], RSTPReid [26], and ICFG-PEDES [27]. CUHK-PEDES integrates 40,206 images of 13,003 individuals from five person search datasets: CUHK03 [28], Market-1501 [29], SSM [30], VIPER [31], and CUHK01 [32]. Each image is annotated with two sentences, totaling 80,412 descriptions. RSTPReid includes 20,505 images of 4101 individuals and is constructed from MSMT17 [33]. Each identity has five images captured by different cameras, with each image paired with two textual descriptions. ICFG-PEDES, also derived from MSMT17, consists of 54,522 images of 4102 individuals, each accompanied by one textual description. Our method is evaluated on the three public text-based person search datasets: CUHK-PEDES, RSTPReid, and ICFG-PEDES.

Evaluation metrics. Following previous works on text-based person search, we adopt the mean Average Precision (AP) and Recall@1,5,10 as our primary evaluation metrics. The Recall@K, whose value is 1 if the first matched image has appeared before the K-th image. Recall@K is sensitive to the position of the first matched image and suits the test set with only one true-matched image in the gallery. The average precision (AP) is the area under the PR (Precision-Recall) curve, considering all ground-truth images in the gallery. mAP is calculated and averaged for the average accuracy (AP) of each category.

Implementation Details Our model is based on the current advanced two-stage benchmark model, and all experiments are trained

using Pytorch on eight NVIDIA A800 GPUs. In pre-training, the Model image encoder uses Swinv2-B as the backbone model [34]. Text encoder and cross encoder use BERT-base [35], respectively. The first 6 and last 6 layers are initialized. At the same time, for the pre-training dataset MALS, we adopt the data filtering method [4] to screen and retrain the image text dataset with a high matching degree for pre-training. We pre-train the model on 32 epochs with a small batch size of 70 per GPU. We use the AdamW [36] optimizer with a weight attenuation of 0.01. In the first 2600 steps learning rate from $1e^{-5}$ begins to warm up, according to the linear plan, and then from $1e^{-4}$ goes down to $1e^{-5}$. Each image input is adjusted to 384×384. Random horizontal inversion, RandAugment [37], and random erase [38] are used for image enhancement. In addition to the image data enhancement mentioned in the pre-training, we also adopt EDA [39] for text data enhancement and set the small batch size to 35. After pre-training, the model is fine-tuned for 30 epochs on the downstream dataset, with a small batch size of 35 per GPU. Set the learning rate to $1e^{-4}$ in the image Encoder, and warm up for the first 3 epochs. Then a linear scheduler is applied to gradually attenuate the learning rate. In the finetune stage, different images and texts with the same ID are randomly selected as weak pairs for training. At the same time, for the MALS dataset used in the pre-training stage, we implement a data filtering strategy [4] to remove the low-quality training data.

Compute/Memory Overhead. Our proposed GITM does not introduce any additional learnable modules or parameters; it only modifies the ITM pair construction by incorporating weak positives and group-wise hard negatives. Under the same training setup described above (same backbone/model configuration, batch size, and optimization settings), GITM (neg3v6) increases the wall-clock training time per epoch from 9:47 to 13:50 and the peak GPU memory from 70776 MiB to 80060 MiB. This overhead is expected since neg3v6 expands the ITM supervision to a group-wise composition with 3 positive pairs (1 strong + 2 weak) and 6 negative pairs, thus evaluating more image–text pairs within the ITM branch. We include this to make the efficiency trade-off of the proposed pairing strategy explicit. **Optimization of γ.** To ensure numerical stability and enforce the positivity constraint, we parameterize the learnable scale γ in log-space and recover it via an exponential mapping. In practice, we optimize a scalar parameter (`log_gamma`) initialized by $\log(1.0)$ and compute $\gamma = \exp(\texttt{log\_gamma})$ during training. This guarantees $\gamma > 0$ throughout optimization and avoids non-positive scaling. Importantly, this computational overhead is strictly limited to the training phase. At inference, the retrieval speed remains unchanged from the baseline, since the proposed method introduces no additional learnable modules and leaves the evaluation pipeline unchanged. Under the same single-GPU A800 evaluation setting, both the baseline and our method showed an inference time of about 9 min, confirming that the proposed training-time modifications do not introduce additional test-time latency.

4.2. Comparison with state-of-the-art methods

On three benchmark datasets, CUHK-PEDES, RSTPReid, and ICFG-PEDES, we compare the proposed method with other advanced text-based person retrieval methods that have reported results or can be re-implemented. The performance evaluation indicators are mean Average Precision (AP) and Recall@1,5,10.

Performance comparison on CUHK-PEDES: We compare our method with lots of competitive methods on CUHK-PEDES. The performance of our method on CUHK-PEDES is shown in Table 2, from which we can observe that:

- (1) Our method achieves a state-of-the-art mAP of 72.44%, along with leading performance in Recall@1, Recall@5, and Recall@10, with scores of 77.88%, 91.05%, and 94.57%, respectively, significantly outperforming other methods. In particular, in terms of mAP, the accuracy of models with uncertainty is improved by +3.06% over RaSa [9] on CUHK-PEDES.

Table 2
Performance comparison on CUHK-PEDES. For a fair comparison, we re-implement APTM [3] with backbone Swinv2-B as Baseline. * indicates the use of additional information, *e.g.*, human parsing.

Method	R@1	R@5	R@10	mAP
Dual Path [14]	44.40	66.26	75.07	–
MIA [40]	53.10	75.00	82.90	–
DSSL [26]	59.98	80.41	87.56	–
SSAN [27]	61.37	80.15	86.73	–
TIPCB [13]	63.63	82.82	89.01	–
LBUL [10]	64.04	82.66	87.22	–
CAIBC [41]	64.43	82.87	88.37	–
LGUR [8]	65.25	83.12	89.00	–
TransTPS [42]	68.23	86.37	91.65	–
CFine [43]	69.57	85.93	91.15	–
VGSG [44]	71.38	86.75	91.86	67.91
MACF [45]	73.33	88.57	93.02	–
IRRA [6]	73.38	89.93	93.71	66.13
TBPS-CLIP [46]	73.54	88.19	92.35	65.38
SAMC [47]	74.03	89.18	93.31	68.42
RDE [48]	75.94	90.14	94.12	67.56
RaSa [9]	76.51	90.29	94.25	69.38
APTM [3]	76.53	90.04	94.15	66.91
ITSELF [49]	76.95	90.64	94.36	69.38
DiCo [50]	77.21	**91.85**	**95.63**	–
AUL [51]	77.23	90.43	94.41	–
BAMG* [52]	79.98	92.31	94.03	68.55
Baseline	76.90	90.75	94.33	68.85
Ours	**77.88**	91.05	94.57	**72.44** (+3.59)

- (2) Comparing to our re-implemented baseline, *i.e.*, APTM + Swinv2-B (Recall@1/5/10: 76.90%, 90.75%, 94.33%, mAP: 68.85%), which adopts conventional contrastive learning and image-text matching objectives, our proposed uncertainty-aware method achieves a +3.59% improvement in mAP and a +0.98% improvement in Recall@1. Notably, the much larger gain in mAP than in Recall@K suggests improved ranking quality beyond top-K hits, *i.e.*, more positive samples are promoted to higher positions throughout the ranked retrieval list, which is better reflected by mAP.
- (3) At the same time, we can observe that the proposed method outperforms the source domain model, *i.e.*, APTM (Recall@1, 5, 10: 76.53%, 90.04%, 94.15%, mAP: 66.91%). This indicates that our uncertainty-based approach effectively leverages a broader range of sample information, leading to a more balanced representation space distribution. By employing group-wise ITM, the model is exposed to a greater diversity of negative samples, significantly contributing to the overall performance improvement.
- (4) The proposed method also surpasses, *i.e.*, RaSa (mAP: 69.38%), which employs relation and sensitivity-aware representation learning. Our uncertainty-aware approach proves more effective in achieving mAP improvements. By employing our proposed uncertainty-based approach, the model can more effectively utilize weak positive pairs to learn a richer feature representation space. This enhancement facilitate the model to learn discriminative feature, allowing the model to correctly identify and rank more positive candidates at higher retrieval hierarchy.

Performance comparison on RSTPReid and ICFG-PEDES: The performance of our model on RSTPReid and ICFG-PEDES is shown in Tables 3 and 4 respectively, and we can observe similar performance improvement: (1) The proposed method is significantly superior to other models, obtaining 69.45% Recall@1, 85.50% Recall@5, 91.65% Recall@10 and 56.11% mAP on RSTPReid. On ICFG-PEDES, 69.22% Recall@1, 83.56% Recall@5, 88.13% Recall@10 and 48.23% of mAP are obtained. In particular, in terms of mAP, the accuracy of models with uncertainty is improved by +3.55% over APTM [3] on RSTPReid and +6.94% over RaSa [9] on ICFG-PEDES. (2) Using the same Swinv2-B backbone, the proposed method achieves competitive results on RSTPReid and ICFG-PEDES, with mAP improvements of +2.89% and

Table 3
Performance comparison on RSTPReid. Baseline: We re-implement APTM [3] with backbone Swinv2-B. * indicates the use of additional information, *e.g.*, humman parsing.

Method	R@1	R@5	R@10	mAP
DSSL [26]	32.43	55.08	63.19	–
LBUL [10]	45.55	68.20	77.85	–
IVT [2]	46.70	70.00	78.80	–
CAIBC [41]	47.35	69.55	79.00	–
CFine [43]	50.55	72.50	81.60	–
TransTPS [42]	56.05	78.65	86.75	–
IRRA [6]	60.20	81.30	88.20	47.17
SAMC [47]	60.80	82.35	89.00	49.67
TBPS-CLIP [46]	61.95	83.55	88.75	48.26
RaSa [9]	66.90	86.50	91.35	52.31
APTM [3]	67.50	85.70	91.45	52.56
RDE [48]	65.35	83.95	89.90	50.88
ITSELF [49]	67.30	85.60	90.50	53.05
DiCo [50]	67.84	85.72	91.98	–
BAMG* [52]	69.73	87.65	93.33	55.21
AUL [51]	**71.65**	**87.55**	**92.05**	–
Baseline	66.75	85.70	91.65	53.22
Ours	69.45	85.50	91.65	**56.11** (+2.89)

Table 4
Performance comparison on ICFG-PEDES. Baseline: We re-implement APTM [3] with backbone Swinv2-B. * indicates the use of additional information, *e.g.*, humman parsing.

Method	R@1	R@5	R@10	mAP
Dual Path [14]	38.99	59.44	68.41	–
MIA [40]	46.49	67.14	75.18	–
SCAN [53]	50.05	69.65	77.21	–
SSAN [27]	54.23	72.63	79.53	–
IVT [2]	56.04	73.60	80.22	–
LGUR [8]	59.02	75.32	81.56	–
CFine [43]	60.83	76.55	82.42	–
MACF [45]	62.95	79.93	85.04	–
IRRA [6]	63.46	80.25	85.82	38.06
SAMC [47]	63.68	79.69	85.21	42.41
TBPS-CLIP [46]	65.05	80.34	85.47	39.83
RaSa [9]	65.28	80.04	85.12	41.29
RDE [48]	67.68	82.47	87.36	40.06
APTM [3]	68.51	82.99	87.56	41.22
DiCo [50]	67.81	83.29	87.62	–
AUL [51]	69.16	83.32	88.37	–
ITSELF [49]	**69.23**	82.84	87.62	43.80
BAMG* [52]	71.70	86.34	89.71	42.37
Baseline	68.71	**83.67**	**88.39**	44.28
Ours	69.22	83.56	88.13	**48.23** (+3.95)

+3.95%, respectively. Additionally, We achieve significant improvements of +2.70% and +0.51% in Recall@1. Our proposed uncertainty-aware ITC and group-wise ITM approach enhances model retrieval capabilities across both datasets by leveraging diverse weak positive samples as a supplement and incorporating more negative samples into ITM learning through a group-wise method. This approach enables the model to retrieve more correctly ranked positive samples, resulting in a significant improvement in mAP performance.

Performance comparison on the Domain Generalization (DG) task. Our method effectively utilizes information from weak positive image-text pairs as supplementary. This approach promotes a more uniform distribution in the model representation space, which naturally suggests that the model can generalize well to other domains. To validate this, we conduct experiments on Domain Generalization (DG) tasks. Specifically, we directly deploy the model, pre-trained on the source domain, to target datasets without further fine-tuning. As shown in Table 5, our method outperforms all other compared approaches. Notably, our method surpasses VGSG [44] by +11.34% in Rank-1 accuracy on the C → I task, and by +22.16% in Rank-1 accuracy on the I → C task. In Table 6, we present the performance of our

Table 5
Comparison results (%) on the domain generalization tasks (*i.e.*, CUHK-PEDES to ICFG-PEDES (C → I) and ICFG-PEDES to CUHK-PEDES (I → C)). The **bold** and underline texts denote the best and runner-up results, respectively.

Method	I → C			C → I		
	R@1	R@5	R@10	R@1	R@5	R@10
Dual Path [14]	15.41	29.80	38.19	7.63	17.14	23.52
MIA [40]	19.35	36.78	46.42	10.93	23.77	32.39
SCAN [53]	21.27	39.26	48.83	13.63	28.61	37.05
SSAN [27]	24.72	43.43	53.01	16.68	33.84	43.00
LGUR [8]	34.25	52.58	60.85	25.44	44.48	54.39
VGSG [44]	35.85	55.04	63.61	27.17	47.77	57.27
Ours	**47.19**	**70.27**	**78.09**	**49.33**	**68.61**	**75.79**

Table 6
Our results (%) on the domain generalization tasks (*i.e.*, CUHK-PEDES to RSTPReid (C → R) and RSTPReid to CUHK-PEDES (R → C), RSTPReid to ICFG-PEDES (R → I) and ICFG-PEDES to RSTPReid (I → R)).

Tasks	Method	R@1	R@5	R@10
C → R	Ours	56.35	77.30	85.30
R → C	Ours	39.49	61.88	70.63
R → I	Ours	45.04	60.71	67.28
I → R	Ours	55.70	74.55	82.45

method on four additional Domain Generalization (DG) tasks. These experiments demonstrate that our uncertainty-aware method exhibits strong generalization capabilities.

To further isolate the contribution of each proposed component under domain shift, we additionally conduct a cross-domain ablation study under the same source-only transfer protocol on the representative I→C and C→I tasks. As shown in Table 8, both uncertainty-aware ITC and GITM remain effective in the cross-domain scenario, and their combination yields the strongest overall transfer performance. On I→C, the full model improves R@1/mAP from 45.96/40.85 to 47.19/44.10. On C→I, it improves R@1/mAP from 47.06/25.60 to 49.33/28.53. These results further support that the DG gains stem from the proposed methodology itself rather than target-side adaptation, since all evaluations are conducted without target-domain fine-tuning.

Discussion. With the added comparisons to recent CLIP-based TBPS systems (e.g., ITSELF, IRRA/RDE, BAMG, DiCo), our method achieves the best mAP performance on CUHK-PEDES and RSTPReid, and ICFG-PEDES while providing a consistent gain over the strong re-implemented baseline across all benchmarks. Notably, our contribution is *pair-level* (uncertainty-aware optimization for noisy/ambiguous correspondences) and is therefore orthogonal to architecture-centric designs (*e.g.*, graph modeling, slot-based disentanglement, fine-grained alignment modules), suggesting potential complementarity when combined.

4.3. Ablation studies and further discussion

To further evaluate our approach, we conduct several ablation studies, with a primary focus on the fine-tuning stage. This emphasis is because our methodology aims to enhance model performance through targeted adjustments to model loss during the fine-tuning phase.

Effect of our uncertainty-aware ITC loss and group-wise ITM loss. We show the ablation comparison of our completely proposed experimental methods in Table 7. (1) First, we filter the dataset based on [4] and conduct experiments based on the initial baseline [3]. (2) Second, the image encoder backbone is replaced by Swinv2-B (the input image size is adjusted to 384 × 384), which shows that the learning ability of the model is further improved. (3) Third, we start to replace the backbone model as the baseline and gradually increase our uncertainty method on it. Firstly, we verify the method of applying uncertainty to adjust ITC loss. It can be seen that the model further

Table 7
Ablation study of our method with different settings on CUHK-PEDES. Baseline†: we re-implement APTM [3] with backbone Swinv2-B for a fair comparison. $\mathcal{L}_{uitc}$ is the optimization objective that uses uncertainty-aware ITC to leverage information about the weak positive pairs fully. $\mathcal{L}_{gitm}$ (neg3v4) is that we adopt the weak positive pairs to increase the number and difficulty of negative pairs by using Group-wise Image-Text Matching (GITM), comprising 1 positive pair, 2 weak positive pairs, and 4 negative pairs. Similarly, $\mathcal{L}_{gitm}$ (neg3v6) is expanded to 1 positive pair, 2 weak positive pairs, and 6 negative pairs.

Method	$\mathcal{L}_{uitc}$	$\mathcal{L}_{gitm}$ (neg3v4)	$\mathcal{L}_{gitm}$ (neg3v6)	R@1	R@5	R@10	mAP
Baseline				76.53	90.04	94.15	66.91
Baseline†				76.90	90.76	94.33	68.86
Baseline	✓			76.88	90.60	94.35	70.49
Baseline	✓	✓		76.85	90.77	94.54	71.89
Baseline	✓		✓	**77.88**	**91.05**	**94.57**	**72.44**

Table 8
Cross-domain ablation results (%) under the source-only transfer protocol on the domain generalization tasks, i.e., ICFG-PEDES → CUHK-PEDES (I → C) and CUHK-PEDES → ICFG-PEDES (C → I). "Baseline† + $\mathcal{L}_{uitc}$" uses only the uncertainty-aware ITC objective, "Baseline† + $\mathcal{L}_{gitm}$" uses only the GITM objective, and "Ours" combines both components. All models are trained on the source domain and directly evaluated on the target domain without target-domain fine-tuning. **Bold** denotes the best result in each column.

Method	I → C				C → I			
	R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP
Baseline†	45.96	68.52	76.51	40.85	47.06	68.13	75.02	25.60
Baseline† + $\mathcal{L}_{uitc}$	45.87	68.67	76.43	41.81	47.66	68.35	75.15	26.39
Baseline† + $\mathcal{L}_{gitm}$ (neg3v6)	45.35	69.33	77.62	41.88	48.44	**68.83**	**75.91**	27.21
Ours	**47.19**	**70.27**	**78.09**	**44.10**	**49.33**	68.61	75.79	**28.53**

increases mAP +1.63% while holding Recall@k. (4) Finally, uncertainty is applied to expand the hard negative in ITM and increase the difficulty of the hard negative. It can be seen that we finally expand to 3 positive pairs and 6 negative pairs. This strategy yields significant improvements in both Recall@k and mAP. Specifically, our method surpasses the baseline (Swinv2-B) on mAP by +3.58% and on Recall@1 by +0.98%.

Comparison of the hard negative number for group-wise image-text matching (GITM). We further evaluate the effect of uncertainty-adjusted hard negative counts in Table 7. $\mathcal{L}gitm$ (neg3v4) expands each example to 3 positive and 4 negative pairs via uncertainty. The set includes the original one-to-one positive pair; two hard negatives obtained by comparing features of the positive pair and selecting the most similar negatives for both image and text; and a new positive formed by pairing the positive image with a weak positive text, with one corresponding negative built from that weak text. Analogously, another positive and its negative are created using a weak positive image. Similarly, $\mathcal{L}gitm$ (neg3v6) constructs 3 positive and 6 negative pairs (details in Section 3.4). We observe that increasing the number of negatives in GITM is more critical. Adding GITM improves performance over not using it: mAP +0.55%, Recall@1 +1.03%, Recall@5 +0.28%, Recall@10 +0.03%.

Analyze the influence of image feature dimension embedding dimension. In Table 9, we explore the influence of the image feature embedding dimension. We observe that increasing the embed dimension from 256 to 2048 leads to improvements across various evaluation metrics, suggesting that a higher embed dimension enhances the model perceptual and learning capabilities. Consequently, all subsequent experiments are conducted with a 2048 embed dimension.

Analyze the influence of loss weight in front of uncertainty-aware ITC loss. We study the impact of the loss weight in front of uncertainty-aware ITC loss. In particular, we change the alpha scale in Eq. (10). We show the effect of different alpha scales on model performance in Table 10. When comparing the impact of alpha on the model, Swin-B is used as the backbone. In order to verify the impact of alpha on uncertainty-aware ITC, we only apply uncertainty learning to regulate ITC loss. Without using uncertainty to increase hard negative, we can observe that the model gets the best performance with uncertainty when $\alpha = 0.5$. At the same time, the experimental results also show that the smaller the proportion of ITC loss adjusted by our uncertainty method in the total loss, the higher the performance of Recall@5, 10, the larger the proportion, and the better the values of mAP and Recall@1, but the higher the proportion is not the better. The experiment shows that the model achieves the best performance when alpha is 0.5.

Table 9
Impact of different ITC embedding dimensions on our model. We report the recall rate and mAP on CUHK-PEDES. Here we only deploy the uncertainty-aware ITC loss. We have achieved the best mAP when the embedding dimension is 2048.

Embedding_dim	R@1	R@5	R@10	mAP
256	76.25	90.10	93.84	68.94
1024	76.48	90.16	93.94	68.83
2048	76.25	90.17	93.91	68.96

Table 10
Impact of different loss weights α of uncertainty-aware ITC on our model. Here we report the recall rate and mAP on CUHK-PEDES. We have achieved the best mAP when the loss weight α of uncertainty-aware ITC is 0.5.

α	R@1	R@5	R@10	mAP
0.3	76.58	**90.34**	**94.23**	70.17
0.4	76.53	89.96	94.07	70.22
0.5	**76.92**	90.11	94.09	**70.78**
0.6	76.59	90.16	94.04	70.73
0.7	76.24	90.15	94.10	70.26

Table 11
Ablation study on the loss weight β. All experiments are conducted on CUHK-PEDES with a fixed $\alpha = 0.1$.

β	R@1	R@5	R@10	mAP
0.01	77.37	90.95	**94.74**	70.23
0.05	77.57	**91.16**	94.41	71.68
0.1	**77.88**	91.05	94.57	72.44
0.2	76.93	90.04	93.63	**72.47**
0.3	76.30	89.18	92.77	72.09
0.4	76.07	88.97	92.49	72.05
0.5	76.18	88.56	92.28	72.06

Impact of GITM Loss Weight β. We further study the sensitivity of our framework to the weighting coefficient β for the GITM loss by

Fig. 4. Visualization of the top 10 person search results on CUHK-PEDES, RSTPReid, and ICFG-PEDES. We present results of two sets of text queries for each of the three datasets in descending order based on the similarity. Images in **green** boxes indicate correct matches, while images in **red** boxes represent incorrect matches. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 12
Impact of different input image sizes. Here we report the recall rate and mAP on CUHK-PEDES. The best mAP is achieved when the input image size is 384 × 384.

H ×W	R@1	R@5	R@10	mAP
256 × 256	76.79	90.68	94.23	71.29
384 × 192	77.01	90.77	94.25	70.97
384 × 384	**77.88**	**91.05**	**94.57**	**72.44**
576 × 192	76.79	90.71	94.41	71.07

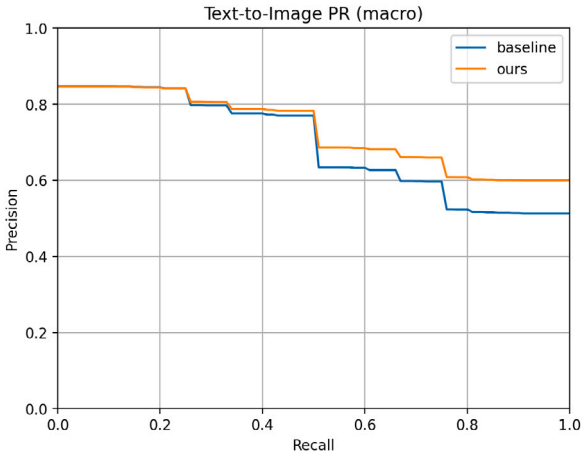
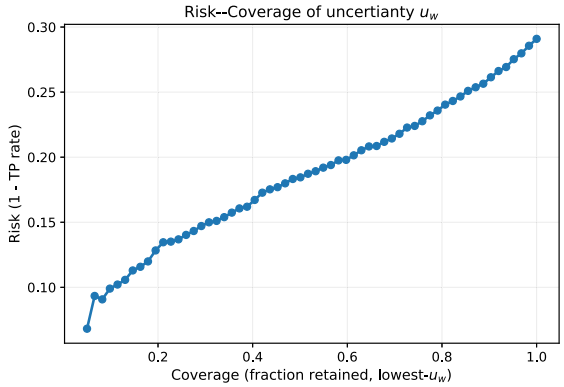


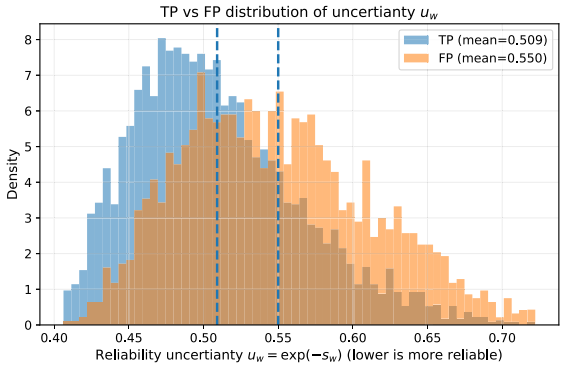
Fig. 5. Macro-averaged Precision–Recall (PR) curves on CUHK-PEDES for text-to-image retrieval. Our method consistently dominates the baseline across recall levels, indicating improved ranking quality beyond top-K recall.

sweeping a wider range from 0.01 to 0.5 on CUHK-PEDES while fixing $\alpha = 0.1$. As shown in Table 11, the performance is relatively robust to β within this range, yet different metrics favor slightly different choices. In particular, $\beta = 0.1$ attains the best R@1 (77.88%) with competitive mAP (72.44%), whereas the highest mAP is achieved at $\beta = 0.2$ (72.47%). For smaller weights, $\beta = 0.05$ yields the best R@5 (91.16%) and a strong mAP (71.68%), and $\beta = 0.01$ gives the best R@10 (94.74%) but a lower mAP (70.23%). When β becomes larger (e.g., ≥ 0.3), Recall@K consistently degrades (from 77.88% R@1 at $\beta = 0.1$ to 76.07% at $\beta = 0.4$), while mAP remains nearly saturated around 72.0%–72.1%. Overall, these results indicate that GITM mainly acts as an auxiliary regularizer: overly increasing its weight does not bring additional benefits and can slightly compromise retrieval recall. Therefore, we adopt $\beta = 0.1$ as the default setting in our experiments.

Comparing the impact of different input sizes in image backbone. We consider the size of the input image and window size to perceive the model receptive field and learning details. All experimental results are obtained after applying all uncertainty methods. For specific results, refer to Table 12. It can be seen from the table that using 384 × 384 as H × W for image processing for the three existing datasets will achieve the best results in the case of Recall@1, 5, 10, and mAP. Additionally, we observe that reducing both the height and width of the images to 256 × 256 leads to a decline in all performance metrics. Similarly, keeping the height constant while reducing the width to 192 results in a performance drop, particularly in mAP. On the other hand, increasing the height to 576 while keeping the width at 192 causes a decrease in R@1 and R@5, but an improvement in R@10 and mAP.

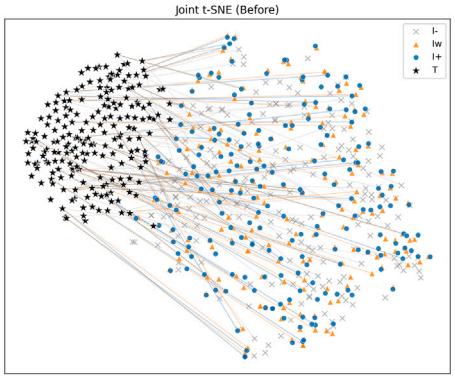


(a) Risk–coverage curve by sorting queries with u_w (lower is more reliable).

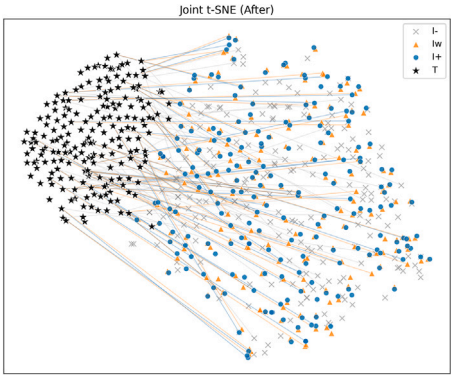


(b) Distribution of u_W for correct vs. incorrect top-1 retrievals.

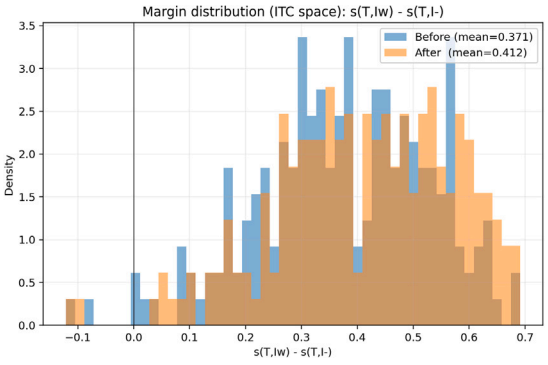
Fig. 6. Reliability analysis of the consistency-based uncertainty. We emphasize that $u_w = \exp(-s_w)$ is a *reliability/ambiguity proxy* derived from weak-pair consistency, rather than a calibrated aleatoric/epistemic uncertainty. On CUHK-PEDES test set (N=6156), incorrect top-1 matches show higher u than correct ones (TP mean 0.509 vs. FP mean 0.550), and sorting by u yields a risk–coverage behavior, indicating that u is monotonic with retrieval risk.



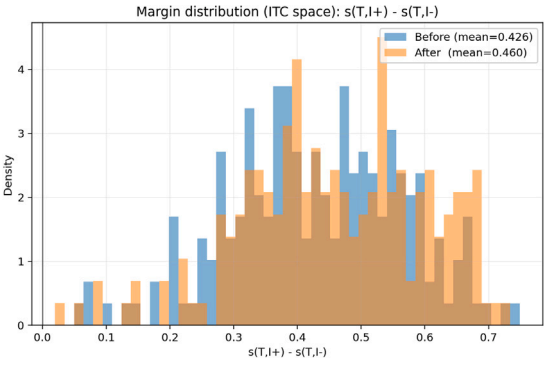
(a) Joint t-SNE (Before).



(b) Joint t-SNE (After).



(c) Weak-positive margin. $s(T, I_w) - s(T, I^-)$.



(d) Positive margin. $s(T, I^+) - s(T, I^-)$.

Fig. 7. Embedding geometry and margin analysis before/after applying our uncertainty-aware learning. (a,b) Joint t-SNE visualization in a shared setting, where $\{T, I^+, I_w, I^-\}$ denote the text query, its matched image, its weak-view counterpart, and a negative image, respectively. Colored connectors indicate the associations from each query to its positives/weak-positives. (c,d) Distributions of ITC margins in the same embedding space. The vertical line at 0 marks the decision boundary where a negative becomes as similar as (or more similar than) the positive/weak-positive. After training, both margin distributions shift right (larger mean margins), indicating more reliable separation against negatives.

Comparison of uncertainty mappings. The uncertainty score in Eq. (3) is instantiated as $u_w = \exp(-s_w)$ in our default setting. Since s_w is computed from cosine similarities on L2-normalized features, it is bounded in $[-1, 1]$, and thus the exponential mapping keeps u_w strictly positive and bounded, which is desirable because u_w appears in the denominator of Eq. (4). To further examine whether this choice is empirically reasonable, we compare it with two simple positive monotonic alternatives, i.e., a linear mapping $u_w = 1.5 - s_w$ and a power-based mapping $u_w = (1.5 - s_w)^2$. As shown in Table 13, the exponential mapping achieves the best overall performance on CUHK-PEDES, outperforming the linear variant by +0.88 R@1 and +0.56 mAP, and also slightly surpassing the power-based variant by +0.15 R@1 and +0.09 mAP. These results suggest that the exponential form provides a more suitable non-linear reweighting for low-consistency

Table 13
Comparison of different uncertainty mappings in Eq. (3) on CUHK-PEDES. All settings are kept identical, and we only replace the mapping from the consistency score s_w to the uncertainty score u_w.

Mapping of u_w	R@1	R@5	R@10	mAP
$1.5 - s_w$	77.00	90.48	94.14	71.88
$(1.5 - s_w)^2$	77.73	90.60	93.96	72.35
$\exp(-s_w)$	**77.88**	**91.05**	**94.57**	**72.44**

Table 14
Effect of introducing pair-level uncertainty into GITM on CUHK-PEDES. "Default" denotes our original design, where u_w is only used in the ITC branch. "Uncertainty-weighted GITM" additionally applies pair-level uncertainty weighting to the weak-positive GITM branches.

GITM supervision	R@1	R@5	R@10	mAP
Default (ours)	**77.88**	**91.05**	**94.57**	**72.44**
Uncertainty-weighted GITM	77.27	90.48	94.19	72.40

Table 15
Effect of removing MALS pre-training on CUHK-PEDES. Here "w/o MALS pre-training" means that we keep the same APTM architecture and standard backbone initialization, but directly train on the downstream dataset without loading the MALS pre-trained checkpoint.

Method	R@1	R@5	R@10	mAP
Baseline w/o MALS pre-training	68.81	86.60	91.42	62.36
Ours w/o MALS pre-training	**70.37**	**87.12**	**91.96**	**66.20**

weak pairs in our current objective. We emphasize that we do not claim $\exp(-s_w)$ to be universally optimal; rather, it is a stable, simple, and empirically effective choice for the present uncertainty-aware ITC formulation.

Does GITM also benefit from uncertainty weighting? To further examine whether u_w should also be introduced into GITM, we implement a variant that applies pair-level uncertainty weighting to the weak-positive GITM branches. As shown in Table 14, this modification does not bring further improvement over the default design: Recall@1/5/10 drop from 77.88/91.05/94.57 to 77.27/90.48/94.19, while mAP changes only marginally from 72.44 to 72.40. This result suggests that uncertainty-aware weighting is more suitable for the ITC branch, where it continuously regulates pair-level ambiguity in the contrastive space, whereas GITM is more effective as an auxiliary binary matching branch whose main gain comes from enriching the group structure and increasing the number and difficulty of hard negatives.

Does our method depend on MALS pre-training? Since our approach is built on the standard pretrain–finetune protocol of APTM, one may ask whether the uncertainty estimation heavily depends on the large-scale MALS pre-trained feature space. To examine this, we conduct an additional experiment without MALS pre-training. Specifically, we keep the same APTM architecture and the standard backbone initialization, but directly train on CUHK-PEDES without loading the MALS pre-trained checkpoint. As shown in Table 15, the baseline achieves 68.81 R@1 and 62.36 mAP, while our full method further improves the performance to 70.37 R@1 and 66.20 mAP. This corresponds to gains of +1.56 R@1 and +3.84 mAP. These results suggest that, although MALS pre-training provides a stronger starting point, the proposed uncertainty-aware learning and GITM do not rely on it to remain effective. The uncertainty signal is still computed online from the current feature space and can provide useful supervision beyond the original MALS-pretrained setting.

4.4. Qualitative results

As shown in Fig. 4, we provide qualitative results of the top 10 search results on three datasets: CUHK-PEDES, RSTPReid, and ICFG-PEDES. Our model uses uncertainty learning to improve precision compared to the baseline. In addition, compared with the conventional baseline approach of contrast learning, we observe that the proposed uncertainty adjustment has better recognition for small-scale targets such as backpacks and tote bags. This is because many descriptions of the same ID have different perspectives, and some perspectives obscure objects such as hand-held objects or backpacks, which makes model learning and text description have limitations. Our uncertainty-aware method corrects these biases and gets reasonable search results.

Ranking-level PR analysis. We further provide a ranking-level diagnostic by plotting the macro-averaged Precision–Recall (PR) curve for text-to-image retrieval on CUHK-PEDES (Fig. 5). For each text query, we rank all gallery images using the final retrieval score (the same score used for mAP evaluation) and define positives as all images sharing the same identity (consistent with our evaluation protocol via `txt2img`). As shown in Fig. 5, our method dominates the baseline across most recall levels, indicating improved ranking quality beyond top-K recall. Quantitatively, Precision@Recall improves from 0.770/0.597/0.514 to 0.783/0.660/0.600 at recall = 0.5/0.7/0.9, and PR-AUC increases from 0.692 to 0.730, with more pronounced gains in the mid-to-high recall regime.

Consistency-based uncertainty reliability analysis. We clarify that our uncertainty is a retrieval-oriented score derived from cross-view (weak-pair) consistency, rather than a probabilistically calibrated aleatoric/epistemic estimate. To validate its reliability meaning, we perform a diagnostic on CUHK-PEDES test set (N=6156). We observe that incorrect top-1 retrievals exhibit higher uncertainty than correct ones (TP mean 0.509 vs. FP mean 0.550), and the resulting risk–coverage curve shows that retaining the lowest-uncertainty fraction substantially reduces error (Fig. 6). These results support that the proposed uncertainty is monotonic with retrieval risk and thus suitable for reliability-aware optimization.

Embedding geometry analysis. To further understand how the proposed uncertainty-aware learning reshapes the joint embedding space, we visualize the representations of sampled tuples $\{T, I^+, I_w, I^-\}$, where I_w denotes a weak-view counterpart under the same identity. As shown in Fig. 7(a–b), compared with the baseline, our method yields a visibly more coherent text–image structure: the associations from T to I^+ and I_w exhibit fewer cross-cluster jumps, suggesting improved cross-view consistency and reduced ambiguity under weak-view perturbations. We complement this qualitative evidence with a margin-based analysis in the ITC space. Fig. 7(c–d) reports the distributions of the margins $s(T, I_w) - s(T, I^-)$ and $s(T, I^+) - s(T, I^-)$, where the 0-line indicates cases where negatives become competitive. After training, both margin distributions shift toward larger values (higher mean margins), indicating that our learning strategy increases the safety margin against negatives not only for the strongest positive pairs (T, I^+) but also for weak-view pairs (T, I_w), which is consistent with the goal of improving reliability under cross-view variations.

5. Conclusion

In this work, we proposed a simple and effective method for text-based person search by explicitly exploiting weak positive pairs that are usually ignored in conventional one-to-one image-text matching. We introduced an uncertainty-aware learning strategy to model the ambiguity of weak positives in cross-modal comparison and incorporated it into the training objective to regularize optimization. We further proposed a group-wise image-text matching scheme to better utilize weak positives and harder negatives, without introducing additional learnable modules or inference-time overhead. Extensive experiments on three benchmark datasets demonstrated the effectiveness of the proposed method, especially in improving mAP, which indicates better overall ranking quality. The results show that weak positive pairs can serve as useful auxiliary supervision when their uncertainty is properly handled. A limitation is that the gain may be reduced when weak pairs are severely mismatched or highly noisy. In future work, we will further explore more robust uncertainty modeling and extend this idea to other related vision-language retrieval tasks.

CRediT authorship contribution statement

Jintao Sun: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Data curation. **Zhedong Zheng:** Writing – review & editing, Project administration. **Gangyi Ding:** Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

We acknowledge supports from Guangdong Basic and Applied Basic Research Foundation 2025A1515012281, Nanjing Municipal Science and Technology Bureau 202401035, and the Macao Science and Technology Development Fund Grant FDCT/0043/2025/RIA1.

Data availability

Data will be made available on request.

References

[1] C. Liu, H. Yang, Q. Zhou, S. Zheng, Making person search enjoy the merits of person re-identification, Pattern Recognit. 127 (2022) 108654, http://dx.doi.org/10.1016/j.patcog.2022.108654.

[2] X. Shu, W. Wen, H. Wu, K. Chen, Y. Song, R. Qiao, B. Ren, X. Wang, 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, Springer-Verlag, Berlin, Heidelberg, 2023, pp. 624–641, http://dx.doi.org/10.1007/978-3-031-25072-9_42.

[3] S. Yang, Y. Zhou, Y. Wang, Y. Wu, L. Zhu, Z. Zheng, Towards unified text-based person retrieval: A large-scale multi-attribute and language search benchmark, in: Proceedings of the 2023 ACM on Multimedia Conference, 2023.

[4] J. Sun, H. Fei, G. Ding, Z. Zheng, From data deluge to data curation: A filtering-wora paradigm for efficient text-based person search, in: ACM WWW, 2025.

[5] X. Ke, H. Liu, P. Xu, X. Lin, W. Guo, Text-based person search via cross-modal alignment learning, Pattern Recognit. 152 (2024) 110481, http://dx.doi.org/10.1016/j.patcog.2024.110481.

[6] D. Jiang, M. Ye, Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval, in: CVPR, 2023.

[7] J. Li, R.R. Selvaraju, A.D. Gotmare, S. Joty, C. Xiong, S.C. Hoi, Align before fuse: vision and language representation learning with momentum distillation, in: NeurIPS, 2021.

[8] Z. Shao, X. Zhang, M. Fang, Z. Lin, J. Wang, C. Ding, Learning granularity-unified representations for text-to-image person re-identification, in: ACM MM, Association for Computing Machinery, New York, NY, USA, 2022, pp. 5566–5574, http://dx.doi.org/10.1145/3503161.3548028.

[9] Y. Bai, M. Cao, D. Gao, Z. Cao, C. Chen, Z. Fan, L. Nie, M. Zhang, RaSa: Relation and sensitivity aware representation learning for text-based person search, in: IJCAI, in: IJCAI-2023, 2023, http://dx.doi.org/10.24963/ijcai.2023/62.

[10] Z. Wang, A. Zhu, J. Xue, X. Wan, C. Liu, T. Wang, Y. Li, Look before you leap: Improving text-based person retrieval by learning a consistent cross-modal common manifold, in: ACM MM, 2022, http://dx.doi.org/10.1145/3503161.3548166.

[11] G. Zhang, Y. Chen, Y. Zheng, G. Martin, R. Wang, Local-enhanced representation for text-based person search, Pattern Recognit. 161 (2025) 111247, http://dx.doi.org/10.1016/j.patcog.2024.111247.

[12] Q. Liu, X. He, Q. Teng, L. Qing, H. Chen, BDNet: A BERT-based dual-path network for text-to-image cross-modal person re-identification, Pattern Recognit. 141 (2023) 109636, http://dx.doi.org/10.1016/j.patcog.2023.109636.

[13] Y. Chen, G. Zhang, Y. Lu, Z. Wang, Y. Zheng, TIPCB: A simple but effective part-based convolutional baseline for text-based person search, Neurocomputing (2022) http://dx.doi.org/10.1016/j.neucom.2022.04.081.

[14] Z. Zheng, L. Zheng, M. Garrett, Y. Yang, M. Xu, Y.-D. Shen, Dual-path convolutional image-text embedding with instance loss, ACM Trans. Multimed. Comput. Commun. Appl. (2020) 1–23, http://dx.doi.org/10.1145/3383184.

[15] A. Kendall, Y. Gal, What uncertainties do we need in Bayesian deep learning for computer vision? in: NeurIPS, NIPS '17, Curran Associates Inc., Red Hook, NY, USA, 2017, pp. 5580–5590.

[16] F. Warburg, M. Jorgensen, J. Civera, S. Hauberg, Bayesian triplet loss: Uncertainty quantification in image retrieval, in: ICCV, 2021, http://dx.doi.org/10.1109/iccv48922.2021.01194.

[17] Y. Chen, Z. Zheng, W. Ji, L. Qu, T.-S. Chua, Composed image retrieval with text feedback via multi-grained uncertainty regularization, in: ICLR, 2024.

[18] J. Postels, M. Segu, T. Sun, L. Van Gool, F. Yu, F. Tombari, On the practicality of deterministic epistemic uncertainty, Int. Conf. Mach. Learn. (2022).

[19] Z. Zheng, Y. Yang, Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation, Int. J. Comput. Vis. 129 (4) (2021) 1106–1120, http://dx.doi.org/10.1007/s11263-020-01395-y.

[20] J. Chang, Z. Lan, C. Cheng, Y. Wei, Data uncertainty learning in face recognition, in: CVPR, 2020, http://dx.doi.org/10.1109/cvpr42600.2020.00575.

[21] Z. Dou, Z. Wang, W. Chen, Y. Li, S. Wang, Reliability-aware prediction via uncertainty learning for person image retrieval, in: S. Avidan, G. Brostow, M. Cissé, G.M. Farinella, T. Hassner (Eds.), ECCV, Springer Nature Switzerland, Cham, 2022, pp. 588–605.

[22] S.J. Oh, K. Murphy, J. Pan, J. Roth, F. Schroff, A. Gallagher, Modeling uncertainty with hedged instance embedding, in: ICLR, 2019.

[23] D.J. Marchette, Bayesian networks and decision graphs, Technometrics 45 (2) (2003) 178–179, http://dx.doi.org/10.1198/tech.2003.s141.

[24] Y. Gal, Z. Ghahramani, Dropout as a bayesian approximation: Representing model uncertainty in deep learning, in: International Conference on Machine Learning, PMLR, 2016, pp. 1050–1059.

[25] S. Li, T. Xiao, H. Li, B. Zhou, D. Yue, X. Wang, Person search with natural language description, in: CVPR, 2017, http://dx.doi.org/10.1109/cvpr.2017.551.

[26] A. Zhu, Z. Wang, Y. Li, X. Wan, J. Jin, T. Wang, F. Hu, G. Hua, DSSL: Deep surroundings-person separation learning for text-based person retrieval, in: ACM Multimedia, MM '21, Association for Computing Machinery, New York, NY, USA, 2021, pp. 209–217, http://dx.doi.org/10.1145/3474085.3475369.

[27] Z. Ding, C. Ding, Z. Shao, D. Tao, Semantically self-aligned network for text-to-image part-aware person re-identification, 2021, arXiv preprint arXiv:2107.12666.

[28] W. Li, R. Zhao, T. Xiao, X. Wang, DeepReID: Deep filter pairing neural network for person re-identification, in: 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, http://dx.doi.org/10.1109/cvpr.2014.27.

[29] L. Zheng, L. Shen, L. Tian, S. Wang, J. Bu, Q. Tian, Person re-identification meets image search, 2015, arXiv preprint arXiv:1502.02171.

[30] T. Xiao, S. Li, B. Wang, L. Lin, X. Wang, End-to-end deep learning for person search, 2, (2) 2016, p. 4, arXiv preprint arXiv:1604.01850.

[31] D. Gray, S. Brennan, H. Tao, Evaluating appearance models for recognition, reacquisition, and tracking, PETS, in: Proc. IEEE International Workshop on Performance Evaluation for Tracking and Surveillance, vol. 3, (5) 2007, pp. 1–7.

[32] W. Li, R. Zhao, X. Wang, Human reidentification with transferred metric learning, in: ACCV, Springer, 2013, pp. 31–44.

[33] L. Wei, S. Zhang, W. Gao, Q. Tian, Person transfer GAN to bridge domain gap for person re-identification, in: CVPR, 2018, http://dx.doi.org/10.1109/cvpr.2018.00016.

[34] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: ICCV, 2021, http://dx.doi.org/10.1109/iccv48922.2021.00986.

[35] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019, pp. 4171–4186.

[36] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, in: ICLR, 2019.

[37] E.D. Cubuk, B. Zoph, J. Shlens, Q.V. Le, Randaugment: Practical automated data augmentation with a reduced search space, in: CVPR Workshop, 2020, http://dx.doi.org/10.1109/cvprw50498.2020.00359.

[38] Z. Zhong, L. Zheng, G. Kang, S. Li, Y. Yang, Random Erasing Data Augmentation, AAAI, 2020, pp. 13001–13008, http://dx.doi.org/10.1609/aaai.v34i07.7000.

[39] J. Wei, K. Zou, EDA: Easy data augmentation techniques for boosting performance on text classification tasks, in: EMNLP-IJCNLP, 2019, http://dx.doi.org/10.18653/v1/d19-1670.

[40] K. Niu, Y. Huang, W. Ouyang, L. Wang, Improving description-based person re-identification by multi-granularity image-text alignments, IEEE Trans. Image Process. 29 (2020) 5542–5556.

[41] Z. Wang, A. Zhu, J. Xue, X. Wan, C. Liu, T. Wang, Y. Li, CAIBC: Capturing all-round information beyond color for text-based person retrieval, in: ACM MM, 2022, http://dx.doi.org/10.1145/3503161.3548057.

[42] L. Bao, L. Wei, W. Zhou, L. Liu, L. Xie, H. Li, Q. Tian, Multi-granularity matching transformer for text-based person search, IEEE Trans. Multimed. 26 (2024) 4281–4293, http://dx.doi.org/10.1109/TMM.2023.3321504.

[43] S. Yan, N. Dong, L. Zhang, J. Tang, CLIP-driven fine-grained text-image person re-identification, IEEE Trans. Image Process. 32 (2023) 6032–6046, http://dx.doi.org/10.1109/TIP.2023.3327924.

[44] S. He, H. Luo, W. Jiang, X. Jiang, H. Ding, VGSG: Vision-guided semantic-group network for text-based person search, IEEE Trans. Image Process. 33 (2024) 163–176, http://dx.doi.org/10.1109/TIP.2023.3337653.

[45] M. Sun, W. Suo, P. Wang, K. Niu, L. Liu, G. Lin, Y. Zhang, Q. Wu, An adaptive correlation filtering method for text-based person search, Int. J. Comput. Vis. 132 (10) (2024) 4440–4455, http://dx.doi.org/10.1007/s11263-024-02094-8.

[46] M. Cao, Y. Bai, Z. Zeng, M. Ye, M. Zhang, An Empirical Study of CLIP for Text-Based Person Search, AAAI, 2024, http://dx.doi.org/10.1609/aaai.v38i1.27801.

[47] Z. Lu, R. Lin, H. Hu, Mind the inconsistent semantics in positive pairs: Semantic aligning and multimodal contrastive learning for text-based pedestrian search, IEEE Trans. Inf. Forensics Secur. 19 (2024) 6409–6424, http://dx.doi.org/10.1109/TIFS.2024.3417251.

[48] Y. Qin, Y. Chen, D. Peng, X. Peng, J.T. Zhou, P. Hu, Noisy-correspondence learning for text-to-image person re-identification, in: CVPR, 2024, pp. 27187–27196, http://dx.doi.org/10.1109/CVPR52733.2024.02568.

[49] T.-H. Nguyen, H.-L. Tran, T.D. Ngo, ITSELF: Attention guided fine-grained alignment for vision–language retrieval, in: WACV, 2026.

[50] G. Kim, C. Eom, DiCo: Disentangled concept representation for text-to-image person re-identification, Neurocomputing (2026) 132885, http://dx.doi.org/10.1016/j.neucom.2026.132885.

[51] S. Li, C. He, X. Xu, F. Shen, Y. Yang, H.T. Shen, Adaptive uncertainty-based learning for text-based person retrieval, AAAI 38 (4) (2024) 3172–3180, http://dx.doi.org/10.1609/aaai.v38i4.28101.

[52] K. Cheng, W. Zou, H. Gu, A. Ouyang, BAMG: Text-based person re-identification via bottlenecks attention and masked graph modeling, in: ACCV, 2024, pp. 1809–1826, http://dx.doi.org/10.1007/978-981-96-0966-6_23.

[53] K.-H. Lee, X. Chen, G. Hua, H. Hu, X. He, Stacked cross attention for image-text matching, in: ECCV, 2018.