Leveraging Phrase-Level Consistency between Images and Texts for Text-Based Person Search

Kota Watanabe[1[0009−0003−5825−1523]] and Yuta Shirakawa[1[0009−0001−1882−9330]]

Toshiba Corporation, 1, Komukai-Toshiba-cho, Saiwai-ku, Kawasaki, 212-8582, Japan
{kota.watanabe.t59,yuta.shirakawa.d48}@mail.toshiba

Abstract. Text-based person search (TBPS) aims to retrieve images of specific individuals from surveillance footage based on free-form text descriptions. This technology enhances monitoring efficiency and reduces operator workload in scenarios such as locating suspicious persons or missing individuals. However, conventional methods consider only the global consistency between the entire text and the image, and even state-of-the-art TBPS approaches still fail to distinguish visually similar individuals with only subtle differences, such as clothing or carried items. To address this limitation, we propose a novel TBPS framework that learns and retrieves based on phrase-level consistency between text and images. During training, phrase-level consistency is evaluated by assigning labels to input text at the phrase level according to the similarity to phrases extracted from ground-truth captions associated with the images. During retrieval, similarity scores computed from query phrases are leveraged to refine the global similarity-based ranking in order to lower the rank of visually similar but incorrect individuals. Comprehensive experiments conducted on two public datasets, RSTPReid and ICFG-PEDES, demonstrate that our method significantly outperforms conventional approaches.

Keywords: Text-based person search · Image retrieval · Vision and language · Deep learning

1 Introduction

Text-based person search (TBPS) is a technology that searches for individuals using free-form text descriptions. It improves the efficiency of monitoring operations and reduces the burden on operators by locating suspicious or missing individuals based on eyewitness testimonies from surveillance camera footage. Compared with person search based on attribute recognition, TBPS is not limited by the prior definition of attributes and has the advantage of incorporating semantically higher-level words and expressions, as well as indicators of certainty and ambiguity[6]. In recent years, with the development of vision-language models (VLMs) such as CLIP[13], BLIP[7] and ALBEF[8], approaches that leverage VLMs pre-trained on large-scale datasets and fine-tune them on relatively small TBPS datasets have become mainstream[2,12,10,1,4].

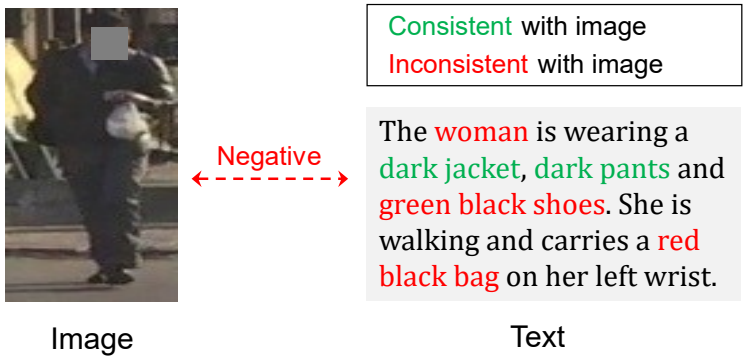

Fig. 1. Example of a negative pair in which the text is partially consistent with the image. Phrases that are consistent with the image are shown in green, while inconsistent phrases are shown in red.

Generally, TBPS datasets are constructed by collecting image-caption pairs describing an individual's clothing and belongings[18,3]. Image-caption pairs linked to the same person are referred to as positive pairs, while pairs linked to different individuals are referred to as negative pairs. The model is trained using classification-based or distance-based learning to determine whether randomly sampled image-caption pairs from the dataset are positive or negative.

Negative pairs constructed from TBPS datasets include phrases in the caption that are partially consistent with the image. For example, the negative pair shown in Figure 1, which consists of an image and a caption, contains phrases such as "dark jacket" and "dark pants" in the caption, and these phrases are consistent with the visual features of the person in the image. Previous TBPS methods have focused solely on determining whether an image–caption pair is positive or negative, and partial consistencies within negative pairs hinder precise model learning.

In this paper, we propose a TBPS method that maximizes the use of textual information through phrase-level consistency, enabling accurate discrimination between the target individual and other visually similar individuals. Our approach consists of two components:

- Training method: This method extracts phrases representing person attributes from the input text, evaluates the consistency of each phrase with the input image based on its ground-truth caption, generates correct labels and uses them for training. By explicitly training the model to predict positive labels for phrases that are consistent with the image (even within negative pairs), this approach mitigates the adverse effect of partial inconsistencies on model learning.
- Re-ranking method: To lower the rank of visually similar individuals who differ only in clothing or belongings, this method calculates a score based on phrase-level similarity in the query text.

2 Related Work

Compared with person retrieval based on attribute recognition[14,17], text-based person search (TBPS) has attracted increasing attention because it is not constrained by predefined attributes and can incorporate higher-level words and expressions, as well as indicators of certainty and ambiguity [6]. TBPS was first proposed by Shuang et al. [9]. In recent years, with the development of large-scale vision-language models (VLMs) such as CLIP [13], BLIP [7], and ALBEF [8], the mainstream approach has been to fine-tune pre-trained VLMs on datasets tailored for TBPS [2,12,10,1,4]. Yang et al. proposed CFine [16], and Min et al. introduced TBPS-CLIP [2], both of which leverage CLIP as the backbone for TBPS. Qin et al. similarly introduced a CLIP-based method called RDE[12]. This method focuses on the prevalence of incorrect image-text correspondences caused by annotation errors and adopts a strategy of preemptively excluding misaligned pairs from the training data based on image-text similarity.

More recently, methods that go beyond simple image-text correspondence and learn from more fine-grained relationships have achieved notable success. Bai et al. proposed RaSa [1], a method built upon ALBEF that incorporates two novel learning strategies: relation-aware learning (RA) and sensitivity-aware learning (SA). RA trains the model to distinguish whether an input positive image-text pair is a strongly related positive pair (comprising an image and its directly descriptive caption) or a weakly related positive pair (comprising an image and a caption describing another image of the same person). By differentiating strong and weak positive pairs, RA mitigates noise introduced by weak pairs in which the image and text are not perfectly consistent due to variations in viewpoint or background. SA randomly replaces tokens in the input text with alternative tokens predicted by a momentum model and requires the model to predict whether each token has been replaced.

Building upon RaSa, Ergasti et al. proposed MARS [4], which introduces an MAE decoder for reconstructing masked regions of an image and an attribute loss (AL) that computes classification errors for individual phrases in the text. In AL-based training, phrases composed of adjective-noun pairs are extracted from the text, and classification errors are computed using the feature representations of tokens corresponding to these phrases from the cross-modal encoder output. Incorporating AL improves discrimination performance by encouraging the model to focus on specific textual phrases. The MAE decoder reconstructs the randomly masked regions in an image using masked auto encoder[5]. However, we argue that AL in MARS still faces the following challenges. When an image–text pair is a negative pair, it is uncertain whether the phrases extracted from the text are actually inconsistent with the image. For example, as illustrated in Figure 1, generic appearance-related phrases such as "dark jacket" often match the image even in negative pairs. Nevertheless, because the image–text pair is labeled as negative, AL trains the model to classify the features of all phrases into the negative class.

3 Proposed Method

In this section, we present the proposed TBPS method. As discussed in Section 2, AL in MARS [4] suffers from the issue that phrase labels do not necessarily remain consistent with the image. To address this problem, our method introduces an improved phrase extraction approach that takes word dependencies into account and trains the model using phrase-level labels generated by our novel labeling strategy. Furthermore, to enhance retrieval accuracy, the retrieval process incorporates phrase-level similarity scores between the image and the query text.

First, we describe the overall workflow of our method, followed by the processing in each component, the loss function, and the retrieval strategy.

3.1 Overview

Figure 2 shows an overview of our TBPS method. Our approach is built upon RaSa [1] and MARS [4]. For simplicity, modules that are common to these methods (such as the MAE decoder and certain loss functions) are omitted from both the explanation and the illustration. Novel modules introduced in our method are highlighted with a light blue background.

When a training pair consisting of an image (hereafter referred to as the probe image, I_{probe}) and a text (hereafter referred to as T_{probe}) is obtained, we first retrieve from the dataset a ground-truth caption corresponding to the person in I_{probe} (hereafter referred to as the positive caption, $T_{positive}$). If multiple captions are available for the person in I_{probe}, we collect up to $N_{positive}$ captions. Next, both T_{probe} and $T_{positive}$ are fed into the phrase extraction block (PEB), which extracts informative phrases for identifying the individual from each text (details in Section 3.2). Then, using the phrase-level labeling block (PLLB), we assign soft labels to the phrases extracted from T_{probe} based on $T_{positive}$, indicating whether each phrase is consistent with I_{probe} (details in Section 3.3). After that, T_{probe} and I_{probe} are input into the text encoder $\mathcal{E}_t$ and image encoder $\mathcal{E}_v$ of ALBEF [8], respectively. The resulting features are then passed to the cross-modal encoder $\mathcal{E}_{cross}$, which re-extracts text features while considering image features. Finally, the output features of the class token and the tokens corresponding to each extracted phrase are fed into two-class classifiers (single fully connected layer), and the classification error is computed against the labels. The label for the class token reflects the relationship between T_{probe} and I_{probe}, while the labels for the phrase tokens are those generated by PLLB. Henceforth, we refer to the loss computed based on the phrase token labels as phrase-level loss ($\mathcal{L}_{PL}$).

During retrieval, we utilize both text-level similarity between the query text and gallery images and phrase-level similarity, enabling clear discrimination between the target individual and other visually similar individuals.

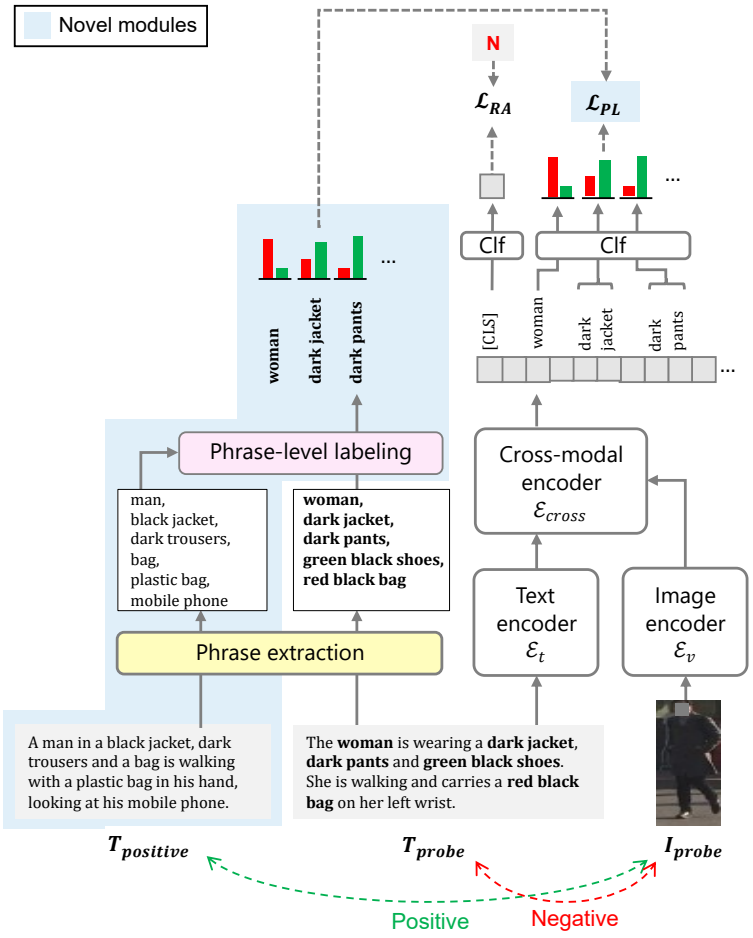

Fig. 2. Overview of our TBPS method. Novel modules are highlighted with a light blue background. For simplicity, the MAE decoder and some loss functions are omitted.

3.2 Improved Method for Extracting Phrases from Text

In the PEB, we extract phrases that help identify individuals from the input text. These phrases, referred to as person attributes, describe characteristics such as gender, age, clothing, and belongings.

To perform phrase extraction, we utilize dependency parse trees. Specifically, we employ the open-source natural language processing library spaCy[1], which enables analysis of both part-of-speech tags and dependency relationships between words in the text. For example, for the sentence "His vest is blue," spaCy produces a dependency parse tree as shown in Figure 3. In this tree, the word "vest" depends on "his" with a possession (poss) relationship, and "is" depends

[1] https://spacy.io/

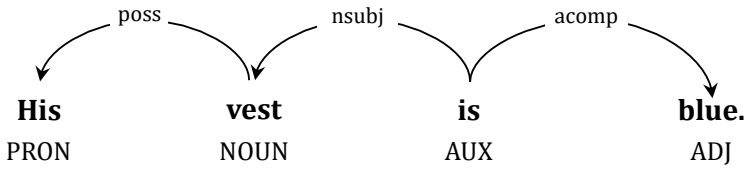

Fig. 3. Example of a dependency parse tree for a text.

on "vest" as its nominal subject (nsubj). Based on the dependency tree, we extract phrases useful for identifying individuals. Starting from nouns in the text, we group together adjectives and nouns that modify them into a single phrase. Since multiple adjectives may be connected, a recursive search is performed on the extracted words. Additionally, if the noun serves as the subject of the sentence, adjectives functioning as complements are also included. For instance, in the sentence "His vest is blue" shown in Figure 3, starting from the noun "vest," the adjective "blue," which functions as a complement, is selected, resulting in the phrase "blue vest." By leveraging dependency structures in this way, our method can extract not only person attributes expressed as adjective–noun sequences (as in MARS [4]) but also those composed of multiple connected nouns and those expressed through subject–predicate relationships. This enables the extraction of a wider variety of person attributes, thereby maximizing the use of informative phrases for identification.

3.3 Method for Generating Labels for Phrases

In the PLLB, we assign labels to determine whether the phrases in T_{probe}, extracted by the PEB described in Section 3.2, are consistent with I_{probe} by referencing the phrases in $T_{positive}$. The procedure for generating phrase-level labels is illustrated in Figure 4.

First, each phrase extracted from T_{probe} is individually fed into the text encoder $\mathcal{E}_t$, and the class token from the output features is obtained. The same process is applied to the phrases extracted from $T_{positive}$. Next, we compute the cosine similarity between each phrase in T_{probe} and each phrase in $T_{positive}$. For every phrase in T_{probe}, the highest cosine similarity with any phrase in $T_{positive}$ is selected. As shown in Figure 4, for the phrase "dark pants" extracted from T_{probe}, the highest cosine similarity is 0.95 with "dark trousers" from $T_{positive}$. This procedure is repeated for all phrases in T_{probe} within the mini-batch. Subsequently, the maximum similarity values for all phrases in the mini-batch are normalized to the range $[0, 1]$ using the minimum and maximum values. Finally, the normalized similarity is used to generate soft labels such that the positive class is assigned the normalized similarity, and the negative class is assigned 1 minus the similarity. These labels are then applied to each phrase.

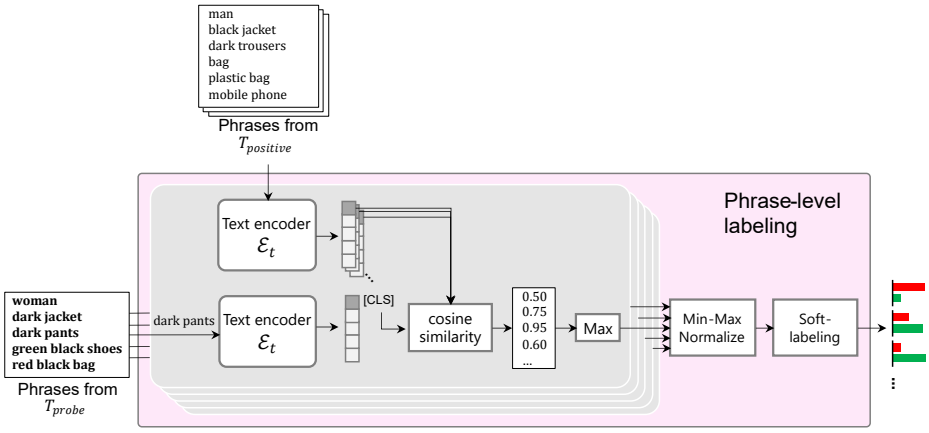

Fig. 4. Detailed processing flow of phrase-level labeling block

3.4 Retrieval Method Considering Phrase-Level Consistency

For the inference procedure that estimates the ranking of gallery images from an input text query, we first outline the conventional approach [1,4] and then describe the differences introduced by our proposed method.

The conventional method first encodes the search query and each image using the text encoder $\mathcal{E}_t$ and image encoder $\mathcal{E}_v$, respectively, and extracts their features. Next, the similarity between the class tokens of the query and image features is computed. For the top k pairs (where $k = 128$) with the highest similarity, the query and image features are further processed by the cross-modal encoder $\mathcal{E}_{cross}$ to refine the features.

Next, the k pairs are re-ranked based on the classification score of the positive class (hereafter referred to as the class score) obtained by feeding the class token features from $\mathcal{E}_{cross}$ into the classifier. In the proposed method, in addition to the class score, we utilize classification scores obtained from individual phrases. Specifically, the features of each phrase are input into the classifier to compute the positive-class score, and the average of these scores (hereafter referred to as the phrase score) is calculated. The final ranking score is obtained by adding the weighted phrase score to the class score. By considering phrase consistency, the ranking of visually similar individuals that differ only in minor attributes can be suppressed, thereby improving retrieval accuracy. The overall score $score_{overall}$ used for re-ranking is defined as follows:

$$score_{overall} = \mathrm{s}_{pos}(x_{cls}) + w_{phr} \cdot score_{phr} \tag{1}$$

$$score_{phr} = \frac{1}{|P|} \sum_{x_{phr} \in P} \mathrm{s}_{pos}(x_{phr})$$

where $score_{cls}$ denotes the class score, $score_{phr}$ denotes the phrase score, x_{cls} represents the class token features, x_{phr} represents the features of each phrase in the set of phrases P, $\mathrm{s}_{pos}(\cdot)$ is a function that computes the positive score from the input features, and w_{phr} is a weight parameter that adjusts the balance between the two scores. We apply robust scaling among the top-k values for each score to mitigate the influence of outliers.

3.5 Loss Function

The loss function used in our method is defined as follows:

$$\mathcal{L} = \underbrace{\mathcal{L}_{p\text{-}ITM} + \lambda_1 \mathcal{L}_{prd}}_{L_{RA}} + \underbrace{\mathcal{L}_{MLM} + \lambda_2 \mathcal{L}_{m\text{-}RTD}}_{L_{SA}} + \lambda_3 \mathcal{L}_{CL} + \lambda_4 \mathcal{L}_{MAE} + \lambda_5 \mathcal{L}_{PL} \tag{2}$$

where $\mathcal{L}_{p\text{-}ITM}$ and $\mathcal{L}_{prd}$ correspond to losses related to relation-aware learning proposed by RaSa [1], while $\mathcal{L}_{MLM}$ and $\mathcal{L}_{m\text{-}RTD}$ correspond to losses related to sensitivity-aware learning. $\mathcal{L}_{CL}$ is the contrastive loss that encourages the class token features output by $\mathcal{E}_t$ and $\mathcal{E}_v$ to be closer for positive pairs and farther apart for negative pairs. Additionally, $\mathcal{L}_{MAE}$ is the reconstruction loss computed by comparing the image reconstructed by the MAE decoder introduced in MARS [4] with the original image. In this experiment, we also introduce $\mathcal{L}_{PL}$ as described in Section 3.1.

To fully exploit phrase-level consistency between images and text, we introduce phrase-level loss ($\mathcal{L}_{PL}$), which is inspired by the concept of AL in MARS [4]. First, from the output of the cross-modal encoder $\mathcal{E}_{cross}$, we obtain token features corresponding to the phrases extracted in Section 3.2. These features are then fed into a classifier to predict a probability distribution. If multiple tokens correspond to a single phrase, we compute the average of their feature vectors. The phrase embedding is calculated as follows:

$$\hat{phr}(T, v, phr) = \frac{1}{N_w^{phr}} \sum_{w \in phr} \mathcal{E}_{cross}(T, v)^{w^i},$$

where N_w^{phr} is the number of words in the phrase, and $\mathcal{E}_{cross}(T, v)^{w^i}$ represents the cross-modal embedding of word w at position i. Here, v is the image embedding obtained from the image encoder $\mathcal{E}_v$.

Next, based on the soft labels (probability distributions) generated in Section 3.3, we compute the error between the predicted distribution and the target distribution using Kullback–Leibler divergence. The loss function is defined as:

$$\mathcal{L}_{PL} = \frac{1}{N_B} \sum_{(I,T) \in S_{it}} \frac{1}{N_{phr}} \sum_{phr \in t} \sum_{c \in C} y_c \log \frac{y_c}{p_c},$$

where p_c is given by

$$p_c = \frac{\exp\left(l_c^{attr}\left(\hat{phr}(T, \mathcal{E}_v(I), phr)\right)\right)}{\sum_{n \in C} \exp\left(l_n^{attr}\left(\hat{phr}(T, \mathcal{E}_v(I), phr)\right)\right)}.$$

where y_c denotes the soft label for class c, p_c represents the predicted probability for class c, N_B is the batch size, S_{it} is the set of image–text pairs in the batch, t denotes the set of phrases extracted from the input text T, N_{phr} is the number of phrases in t, and C is the set of classes (negative/positive). The function $l_c^{attr}(\cdot)$ is the attribute classifier logit for class c applied to a phrase embedding.

4 Experiments

4.1 Datasets

RSTPReid [18] and ICFG-PEDES [3] are used for accuracy evaluation. RSTPReid consists of 20,505 images, each accompanied by two captions. The images are collected from 4,101 individuals, with five images per individual. The dataset is divided into 3,701 individuals for training, 200 for validation, and 200 for testing. ICFG-PEDES consists of 54,522 images, each accompanied by one caption. The images are collected from 4,102 individuals, with an average of 13 images per individual. The dataset is divided into 3,102 individuals for training and 1,000 for testing. Both datasets collect images from MSMT17 [15], and many images are shared between them. Figure 5 shows examples of captions assigned to the same images in both datasets. The images in the upper and lower rows depict the same person. Captions in RSTPReid are short and use similar phrases for the same attributes, whereas captions in ICFG-PEDES are longer and often exhibit paraphrasing, such as describing the same attribute as "medium length black hair" and "short black hair."

4.2 Experimental Setup

The model is trained for 30 epochs for experiments on RSTPReid [18], and for 60 epochs for experiments on ICFG-PEDES [3]. To ensure a fair comparison, RaSa [1] and MARS [4] are trained using the same number of epochs. AdamW [11] is employed as the optimizer with a weight decay of 0.02. The learning rate is initialized at 10^{-5} and decays to a minimum value of 10^{-6} using a cosine scheduler. The weights of each loss term in Equation 2 are set to $\lambda_1 = 0.5$, $\lambda_2 = 0.5$, $\lambda_3 = 0.5$, and $\lambda_4 = 1.0$, respectively. For λ_5, it is set to 3.0 for RSTPReid and 6.0 for ICFG-PEDES. Additionally, $N_{positive}$ is set to 5.

4.3 Evaluation Metrics

We use Rank-N ($N = 1, 5, 10$) and mean average precision (mAP) as evaluation metrics. Rank-N represents the proportion of queries for which at least one

Image	Caption	
	RSTPReid	ICFG-PEDES
	The man with glasses is wearing a grey jacket with a blue shirt inside. He has his hands in his pockets.	A middle-aged man with medium length black hair parted sideways is wearing a dark grey fleece coat with black buttons over blue white striped polo. He has a black bag strap around his shoulders.
	The man is a strong man who is in a grey jacket. His shirt is blue and his trousers are black.	A man in his mid-forties with short black hair is wearing a long grey pea coat, light blue shirt inside the coat over black jeans. He is carrying a backpack with black straps visible on the coat and wearing brown formal shoes.

Fig. 5. Examples of images and captions included in the evaluation datasets.

correct image is included within the top-N predicted images. mAP indicates how well all images considered correct are retrieved at higher ranks. We emphasize Rank-N for two key reasons:

- Due to factors such as the orientation of the person and low image quality, some images labeled as correct do not sufficiently reflect the descriptions in the query text. mAP is highly affected by such noisy images.
- In real-world monitoring tasks, operators need to quickly identify individuals by checking only a few top-ranked images. Therefore, Rank-N is important as it evaluates how highly the top correct image is ranked.

4.4 Results and Analysis

The performance comparison results on RSTPReid are presented in Table 1. Our method outperformed conventional approaches in terms of both Rank-N and mAP metrics. In particular, Rank-1 reached 69.05%, significantly surpassing state-of-the-art methods such as CADA [10] and MARS [4]. By computing scores based on phrase-level consistency, our method demonstrated strong capability in retrieving the correct image at the top ranks when perfectly matching the text description, contributing to improved Rank-1. The performance comparison results on ICFG-PEDES are shown in Table 2. Similar to RSTPReid, our method achieved 67.83% Rank-1, outperforming conventional approaches, and demonstrating comparable performance in other metrics. As illustrated in Figure 5, the captions in ICFG-PEDES contain phrases that are frequently paraphrased in diverse ways, making it difficult to achieve high similarity between text features. Consequently, even phrases that align well with the image may fail to receive high positive labels. Further improvements are expected by refining the labeling strategy, for example by incorporating adjustments based on expression complexity. Moreover, since RDE [12] has demonstrated strong

Table 1. Performance comparison experiments using **RSTPReid**. The best results for each metric are shown in **bold**, and the second-best results are underlined.
* model trained with the same number of epochs as ours for a fair comparison.

Method	Backbone	Rank-1(%)	Rank-5(%)	Rank-10(%)	mAP(%)
—	ALBEF[8]	50.10	73.70	82.10	41.73
—	CLIP[13]	54.05	80.70	88.00	43.41
CFine[16]	CLIP	50.55	72.50	81.60	—
TBPS-CLIP[2]	CLIP	61.95	83.55	88.75	48.26
RDE[12]	CLIP	65.35	83.95	89.90	50.88
CADA[10]	BLIP[7]	67.70	84.60	89.75	49.95
RaSa*[1]	ALBEF	66.60	85.85	91.05	50.98
MARS*[4]	ALBEF	67.05	85.00	90.65	52.91
Ours	ALBEF	**69.05**	**85.90**	**91.70**	**53.63**

Table 2. Performance comparison experiments using **ICFG-PEDES**. The best results for each metric are shown in **bold**, and the second-best results are underlined.
* model trained with the same number of epochs as ours for a fair comparison.

Method	Backbone	Rank-1(%)	Rank-5(%)	Rank-10(%)	mAP(%)
—	ALBEF[8]	34.46	52.32	60.40	19.62
—	BLIP[7]	56.16	73.77	80.17	31.59
—	CLIP[13]	56.74	75.72	82.26	31.84
CFine[16]	CLIP	60.83	76.55	82.42	—
TBPS-CLIP[2]	CLIP	65.05	80.34	85.47	39.83
RDE[12]	CLIP	67.68	**82.47**	**87.36**	40.06
CADA[10]	BLIP	67.38	81.34	85.64	37.81
RaSa*[1]	ALBEF	65.92	81.06	85.88	42.91
MARS*[4]	ALBEF	67.60	81.62	85.97	**44.82**
Ours	ALBEF	**67.83**	81.43	85.98	44.74

performance in the ICFG-PEDES setting, it suggests that, when using captions that are long and information-rich, strategies that minimize incorrect correspondences between images and texts and focus on pairs that are relatively easier to classify could also be effective.

The results of the ablation study on the loss function $\mathcal{L}_{PL}$ and the phrase score $score_{phr}$ used for re-ranking in the proposed method are shown in Table 3. For both datasets RSTPReid and ICFG-PEDES, accuracy was improved by introducing $\mathcal{L}_{PL}$, and further improved by incorporating $score_{phr}$. The improvement was more significant for RSTPReid, suggesting that the proposed method is more effective when text descriptions are concise. Figure 6 shows how Rank-1 and mAP vary as the weight w_{phr} applied to $score_{phr}$ varies. Both Rank-1 and mAP improve as w_{phr} increases from zero, although there is a slight gap in the timing of their respective peaks.

The qualitative comparison of retrieval performance between our method and conventional approaches is shown in Figure 7. For RSTPReid, we present the top-10 retrieved images using the baseline methods MARS [4], RaSa [1], and

Table 3. Ablation study on $\mathcal{L}_{PL}$ and $score_{phr}$. Numbers in parentheses indicate improvements over the baseline.

		RSTPReid		ICFG-PEDES	
$\mathcal{L}_{PL}$	$score_{phr}$	Rank-1 (%)	mAP (%)	Rank-1 (%)	mAP (%)
		67.35	53.05	67.01	43.53
✓		68.15 (+0.80)	53.21 (+0.16)	67.77 (+0.76)	44.68 (+1.15)
✓	✓	69.05 (+1.70)	53.63 (+0.58)	67.83 (+0.82)	44.74 (+1.21)

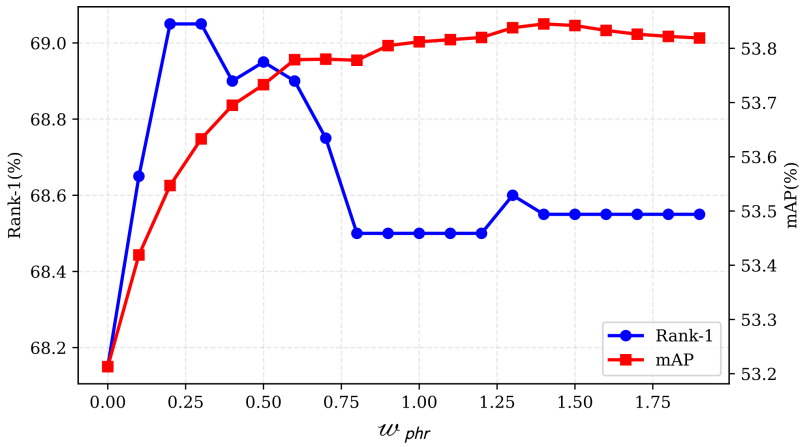

Fig. 6. Changes in Rank-1 and mAP as w_{phr} varies

our proposed method. The correct images corresponding to the query text are highlighted with green bounding boxes. Our method, which ranks results based on phrase-level consistency, effectively assigns lower ranks to individuals who do not fully match the query phrases and higher ranks to those who are consistent with all phrases, demonstrating its effectiveness.

Furthermore, as one possible application of our method, we examine a retrieval approach that emphasizes specific phrases in the query text. This is useful in scenarios where the confidence level for each attribute needs to be adjusted, such as when there are specific clues that should be emphasized during identification based on eyewitness testimony in video footage. Figure 8 shows the retrieval results when the weight of a specific phrase ("suitcase" in the figure) is set significantly higher ($\times 10$ in the figure) than other phrases during the reranking process. Orange frames indicate individuals actually carrying a suitcase. It can be observed that individuals exhibiting the emphasized feature are more frequently retrieved among the top-ranked images. These findings indicate that our method offers highly practical functionality for real-world applications.

Fig. 7. Top 10 search results obtained by our proposed method and conventional methods. The correct person's images are indicated with a green frame.

5 Conclusion and Limitations

In this paper, we proposed a novel TBPS framework that explicitly leverages phrase-level consistency between images and texts. For training, we assign labels to individual phrases in the input text and introduce a loss function that evaluates phrase-level consistency. For retrieval, we proposed a re-ranking method that utilizes similarity scores estimated for each phrase in the query text. Extensive experiments on public benchmarks demonstrate that our approach consistently outperforms conventional methods, particularly in terms of Rank-1 accuracy, which is critical for practical monitoring scenarios. Moreover, we showed that emphasizing specific phrases in the query text enables flexible and interpretable retrieval behavior, highlighting the practical usefulness of the proposed framework.

Despite its effectiveness, this work has several limitations. First, the phrase-level re-ranking stage introduces additional inference cost, as it requires cross-modal encoding of top-ranked candidates and phrase-wise classification. While this design allows fine-grained consistency modeling, it is inherently more computationally demanding than global-only matching. In practical deployments, it is possible to limit the number of phrases used for re-ranking or to apply the

Fig. 8. Effect of emphasizing a specific phrase. Top: Retrieval results using our method without emphasis. Bottom: Retrieval results with emphasis applied. Orange frames indicate individuals actually carrying a suitcase.

re-ranking stage as an offline process; however, improving inference efficiency remains an important direction for future work. Second, the phrase-level labeling relies solely on text-to-text similarity, which makes the method sensitive to paraphrasing. As observed on datasets with diverse linguistic expressions, such as ICFG-PEDES, visually correct phrases may receive lower confidence scores when lexical variations are large. Incorporating image-aware semantic alignment or more robust language normalization is a promising avenue for future research.

Looking ahead, we believe that the concept of phrase-level consistency modeling can be extended beyond TBPS. In general object retrieval scenarios, partial visual matches in negative pairs may similarly hinder effective learning. Likewise, in video-based action recognition, different actions often share overlapping temporal segments. Decomposing labels and explicitly modeling partial consistency at a finer granularity could therefore offer benefits across a broader range of vision and language tasks.

References

1. Bai, Y., Cao, M., Gao, D., Cao, Z., Chen, C., Fan, Z., Nie, L., Zhang, M.: RaSa: Relation and sensitivity aware representation learning for text-based person search. In: Proc. Thirty-Second Int'l Joint Conf. Artificial Intelligence. p. 555–563 (2023)
2. Cao, M., Bai, Y., Zeng, Z., Ye, M., Zhang, M.: An empirical study of clip for text-based person search. In: Proc. the AAAI Conference on Artificial Intelligence. vol. 38, pp. 465–473 (2024)
3. Ding, Z., Ding, C., Shao, Z., Tao, D.: Semantically self-aligned network for text-to-image part-aware person re-identification. arXiv preprint arXiv:2107.12666 (2021)

4. Ergasti, A., Fontanini, T., Ferrari, C., Bertozzi, M., Prati, A.: Mars: Paying more attention to visual attributes for text-based person search. ACM Trans. Multimedia Computing, Communications, and Applications pp. 1–22 (2025)
5. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition. pp. 16000–16009 (2022)
6. Jiang, F., Yang, S., Jones, M.W., Zhang, L.: From attributes to natural language: A survey and foresight on text-based person re-identification. Information Fusion **118**, 102879 (2025)
7. Li, J., Li, D., Xiong, C., Hoi, S.: BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., Sabato, S. (eds.) Proc. 39th Int'l Conf. on Machine Learning. pp. 12888–12900 (2022)
8. Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., Hoi, S.C.H.: Align before fuse: Vision and language representation learning with momentum distillation. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems. vol. 34, pp. 9694–9705 (2021)
9. Li, S., Xiao, T., Li, H., Zhou, B., Yue, D., Wang, X.: Person search with natural language description. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition. pp. 1970–1979 (2017)
10. Lin, D., Peng, Y.X., Meng, J., Zheng, W.S.: Cross-modal adaptive dual association for text-to-image person retrieval. IEEE Trans. on Multimedia **26**, 6609–6620 (2024)
11. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
12. Qin, Y., Chen, Y., Peng, D., Peng, X., Zhou, J.T., Hu, P.: Noisy-correspondence learning for text-to-image person re-identification. In: 2024 IEEE/CVF Conf. Computer Vision and Pattern Recognition. pp. 27187–27196 (2024)
13. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Meila, M., Zhang, T. (eds.) Proc. 38th Int'l Conf. Machine Learning. vol. 139, pp. 8748–8763 (18–24 Jul 2021)
14. Sarafianos, N., Xu, X., Kakadiaris, I.A.: Deep imbalanced attribute classification using visual attention aggregation. In: Proc. European Conference on Computer Vision. pp. 680–697 (2018)
15. Wei, L., Zhang, S., Gao, W., Tian, Q.: Person transfer gan to bridge domain gap for person re-identification. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition. pp. 79–88 (2018)
16. Yan, S., Dong, N., Zhang, L., Tang, J.: Clip-driven fine-grained text-image person re-identification. IEEE Transactions on Image Processing **32**, 6032–6046 (2023)
17. Yang, J., Fan, J., Wang, Y., Wang, Y., Gan, W., Liu, L., Wu, W.: Hierarchical feature embedding for attribute recognition. In: Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition. pp. 13055–13064 (2020)
18. Zhu, A., Wang, Z., Li, Y., Wan, X., Jin, J., Wang, T., Hu, F., Hua, G.: DSSL: Deep surroundings-person separation learning for text-based person retrieval. In: Proc. 29th ACM Int'l Conf. Multimedia. p. 209–217 (2021)