Visual Perturbation for Text-Based Person Search

Pengcheng Zhang[1], **Xiaohan Yu**[2], **Xiao Bai**[1*], **Jin Zheng**[1]

[1]School of Computer Science and Engineering, State Key Laboratory of Complex & Critical Software Environment, Jiangxi Research Institute, Beihang University, Beijing, China
[2]School of Computing, Macquarie University, Sydney, Australia
{pengchengz, baixiao, jinzheng}@buaa.edu.cn, xiaohan.yu@mq.edu.au

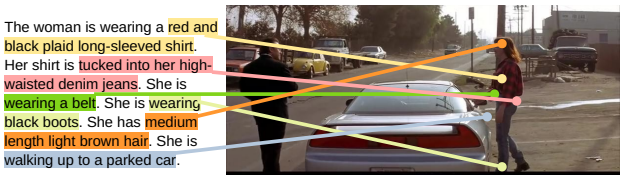
Abstract

Text-based person search aims at locating a person described by natural language in uncropped scene images. Recent works for TBPS mainly focus on aligning multi-granularity vision and language representations, neglecting a key discrepancy between training and inference where the former learns to unify vision and language features where the visual side covers all clues described by language, yet the latter matches image-text pairs where the images may capture only part of the described clues due to perturbations such as occlusions, background clutters and misaligned boundaries. To alleviate this issue, we present ViPer: a Visual Perturbation network that learns to match language descriptions with perturbed visual clues. On top of a CLIP-driven baseline, we design three visual perturbation modules: (1) Spatial ViPer that varies person proposals and produces visual features with misaligned boundaries, (2) Attentive ViPer that estimates visual attention on the fly and manipulates attentive visual tokens within a proposal to produce global features under visual perturbations, and (3) Fine-grained ViPer that learns to recover masked visual clues from detailed language descriptions to encourage matching language features with perturbed visual features at the fine granularity. This overall framework thus simulates real-world scenarios at the training stage to minimize the discrepancy and improve the generalization ability of the model. Experimental results demonstrate that the proposed method clearly surpasses previous TBPS methods on the PRW-TBPS and CUHK-SYSU-TBPS datasets.

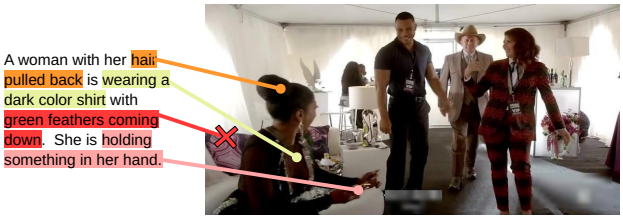
Code — https://github.com/PatrickZad/ViPer

Introduction

Person search aims to recognize and locate a target person among a gallery of raw scene images captured by different cameras. Existing person search tasks can be divided into two categories, *i.e.* image-based person search (IBPS) (Zheng et al. 2017; Xiao et al. 2017) that represents the query persons by their images, and text-based person search (TBPS) (Zhang et al. 2023) that presents only language descriptions of the queries. Despite the great progress in IBPS, those methods are not applicable when the image of a target person is unavailable. In many application scenarios, the language descriptions of a target person can be easier to obtain than prior images. Therefore, exploring how to retrieve persons in scene images based on text descriptions is an important step in extending the capability of person search techniques for real-world demands.



(a) A image-text pair for training.



(b) A positive image-text pair for inference.

Figure 1: Illustration of the discrepancy between training and inference for TBPS. We use colored lines and backgrounds to highlight the matched vision and language clues.

To facilitate the development of TBPS, SDPG (Zhang et al. 2023) proposes two TBPS datasets and designs a vision-language alignment framework across global and fine-grained person features. MACA (Su et al. 2024) jointly performs global-level and attribute-level vision-language alignment to learn unified multi-modal person features. The text-based person ReID (Li et al. 2017; Ding et al. 2021; Zhu et al. 2021) task is also closely related to TBPS. For text-based person ReID, recent works mainly attempt to incorporate pre-trained vision-language models (Jiang and Ye 2023; Cao et al. 2024; Yan et al. 2023) and encourage fine-grained vision-language matching (Yang et al. 2024; Suo et al. 2022; Jiang and Ye 2023; Zuo et al. 2024) implicitly or explicitly. Despite their advances, those works focus on discovering and aligning multi-granularity vision and language features, neglecting a key discrepancy between training, where the

*Corresponding author.

image covers all language-described clues, and inference, where the positive image may match only partial text information. As Figure 1 shows, the language-described clues can be completely found in the paired image during training. During inference, the language description can be as clear and detailed as possible, while the scene images may capture only part of the clues due to perturbations such as occlusions, background clutters, and misaligned boundaries. This hinders the generalization capability of the model to retrieve targets in real-world scenarios.

To minimize the discrepancy between training and inference for TBPS, we propose a **Vi**sual **Per**turbation network (ViPer) in this work. On top of a CLIP-driven (Radford et al. 2021) end-to-end baseline, we design three visual perturbation modules to force vision-language alignment under the condition that the visual features encode only partial language-described clues of the same person. Specifically, we design a spatial perturbation (Spatial ViPer) that varies the person proposals to produce visual features with misaligned boundaries. On top of that, we introduce visual perturbations to both global and fine-grained visual feature extractions prior to aligning the vision and language features. For global visual features, we design attentive perturbations (Attentive ViPer) to adaptively remove highly attended visual tokens or exchange the less attended ones before aggregating them to the global feature. The visual attentions on the tokens are estimated without extra models or annotations to guarantee the computational efficiency. For fine-grained visual features, we design masked visual token modeling to introduce fine-grained perturbation (Fine-grained ViPer). Before projecting the visual tokens into the unified feature space, we randomly replace a ratio of the tokens with a learned mask token and employ a cross-modal transformer to gradually recover the masked tokens. To recover the masked visual tokens only from the language clues, we further propose to employ a full cross-attention architecture for the cross-modal transformer. This implicitly encourages fine-grained vision-language alignment under visual perturbation. By doing so, the overall ViPer simulates the inconsistency between paired image and text information during training to improve the generalization capability of the TBPS model.

To summarize, this work makes the following contributions:

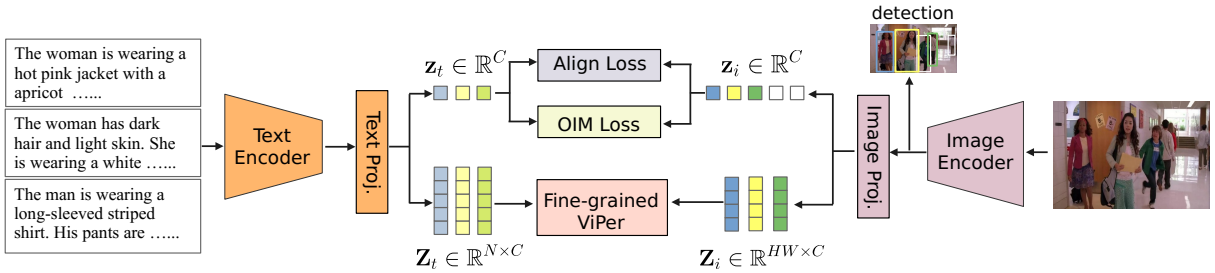
- We propose an end-to-end Visual Perturbation network (ViPer) for TBPS. ViPer minimizes the discrepancy between training and inference image-text pairs to improve the generalization capability of the TBPS model.
- We design three ViPer modules, *i.e.* Spatial ViPer, Attentive ViPer and Fine-grained ViPer. These modules effectively simulate the main visual perturbations that occur in inference for training, facilitating vision-language alignment under visual perturbations at both global and fine-grained scales.
- Experimental results demonstrate that the proposed method achieves superior performances on both the PRW-TBPS and the CUHK-SYSU-TBPS datasets.

Related Works

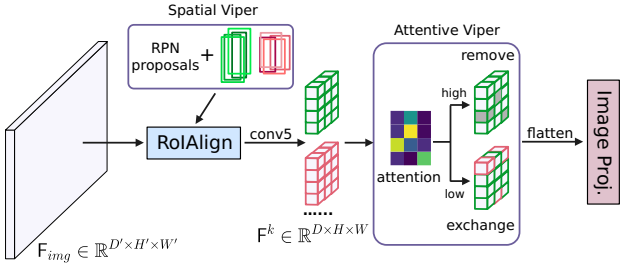
Person Search research starts from the IBPS problem. Existing methods are typically categorized into two types: two-step methods and end-to-end methods. Two-step methods first employ a standalone detector to detect persons in the scene image and then crop the person images. Afterward, an independent ReID model is utilized to retrieve a target person across the cropped images. For two-step person search, Zheng et al. (Zheng et al. 2017) first explored to combine popular person detectors and ReID models. Lan, Zhu, and Gong (Lan, Zhu, and Gong 2018) performed a multi-scale matching between persons to deal with large-scale variances. To reduce distractors during person retrieval, Wang et al. (Wang et al. 2020) designed a target-guided person detector to detect persons in the gallery images conditioned on the queries. To improve the efficiency of the two-step paradigms, end-to-end methods are proposed to perform person search with a unified model. The first end-to-end model is presented by Xiao et al. (Xiao et al. 2017). Following works (Chen et al. 2020b; Han et al. 2021; Yan et al. 2021) then tackled the conflicting subtasks by well-designed training objectives or model architectures. Li and Miao (Li and Miao 2021) further improved the person search performance by refining person proposals. Cao et al. (Cao et al. 2022) and Yu et al. (Yu et al. 2022) designed effective person search transformers to boost the performance.

For TBPS, SDPG (Zhang et al. 2023) constructed two benchmarks and proposed an effective semantic-driven proposal generation model with cross-scale vision-language matching. MACA (Su et al. 2024) mainly designs fine-grained feature discovering and aligning methods for TBPS. Different from these works, this paper focuses on minimizing the discrepancy between training and inference to improve the generalization of the TBPS model.

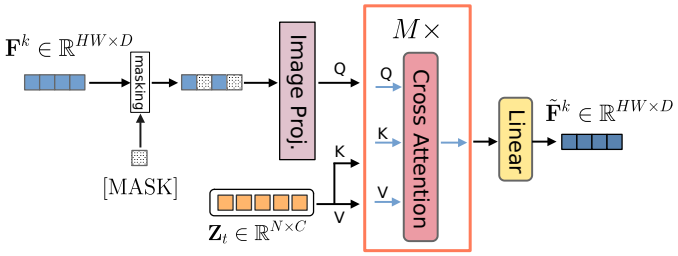
Text-based Person ReID performs text-to-image person retrieval where the images are cropped to bound the person bodies (Li et al. 2017; Ding et al. 2021; Zhu et al. 2021). We also note that recent works employ different terminologies to refer to this task while this work uses 'text-based person ReID' for consistency. To tackle this problem, previous works mainly focus on aligning discriminative global and fine-grained vision and language features. Suo et al. (Suo et al. 2022) designed filtering modules to extract fine-grained key clues and adaptively align the cross-modal discriminative features. Jiang and Ye (Jiang and Ye 2023) implicitly matched local visual-textual tokens via cross-modal masked language modeling (MLM) (Devlin et al. 2018) and proposed improved global image-text aligning loss. Yang et al. (Yang et al. 2024) designed a disordered fine-grained feature learning module to enhance model robustness. Zuo et al. further proposed to employ ultra fine-grained text descriptions for the retrieval task. As the vision language models (VLMs) (Radford et al. 2021; Li et al. 2022b,c, 2023) show advanced capability in vision-language understanding, recent works (Yan et al. 2023; Jiang and Ye 2023; Cao et al. 2024) also explore incorporating the pre-trained VLMs to narrow the gap between the vision and language modalities.



(a) The overall architecture of ViPer.



(b) Illustration of Spatial Viper and Attentive Viper.



(c) Illustration of Fine-grained Viper.

Figure 2: An overview of the proposed method. (a) The overall ViPer is built upon modality-specific encoders and projectors to produce vision and language features. The detection modules are integrated with the image side to enable end-to-end person search. Both the global features ($\mathbf{z}_t$ and $\mathbf{z}_i$) and fine-grained features ($\mathbf{Z}_t$ and $\mathbf{Z}_i$) are optimized with designed visual perturbations for TBPS. (b) Spatial Viper simulates misaligned boundaries in person location proposals, affecting both $\mathbf{z}_i$ and $\mathbf{Z}_i$. Attentive Viper simulates occlusions and background clutters by manipulating tokens before the image projector according to attention on the fly, producing visually perturbed global features. (c) Fine-grained ViPer performs random masking of visual tokens and restores them from the text tokens via cross-attention, facilitating matching fine-grained vision-language features under visual perturbations.

Method

In this section, we first briefly introduce the baseline person search network on which the proposed ViPer is built. We then present the designed Spatial ViPer, Attentive ViPer, and Fine-grained ViPer that form the overall ViPer for TBPS. The details of training and inference are given at last.

Baseline

Following recent works in text-based person ReID (Jiang and Ye 2023; Cao et al. 2024), we base the proposed model on CLIP (Radford et al. 2021) to exploit pretrained vision-language models for TBPS. As illustrated in Figure 2a, we use the image and text encoders, including their correlated projectors, pretrained by CLIP to form an end-to-end TBPS network (Xiao et al. 2017). Due to the computation complexity of processing high-resolution scene images with the transformer (Vaswani et al. 2017; Dosovitskiy et al. 2020) image encoder, we employ the CLIP modified ResNet50 (He et al. 2016) image encoder and the paired text encoder. Similar to previous IBPS methods (Chen et al. 2020a,b; Li and Miao 2021), the image encoder inherits the Faster R-CNN (Ren et al. 2015) pipeline that encodes image features with the first four stages ('conv1-4') and outputs instance features with the last stage ('conv5') upon region proposals. The instance features are used to predict person locations and produce appearance features for retrieval. As the detection subtask is only defined in the visual modality, we directly predict person detection results from the instance features. And the image projector only serves to produce appearance features for person retrieval.

For training of the baseline, we extract and optimize global vision and language person features $\mathbf{z}_t \in \mathbb{R}^C$ and $\mathbf{z}_i \in \mathbb{R}^C$ by an identity classification loss and a vision-language align loss similar to recent text-based ReID models (Jiang and Ye 2023; Cao et al. 2024). Respectively, we employ the widely used Online-Instance-Matching loss (Xiao et al. 2017) $\mathcal{L}_{oim}$ and the Cross-Modal Projection Matching (Zhang and Lu 2018) loss $\mathcal{L}_{cmpm}$. To enable person localization, the model is jointly supervised by the detection losses in Faster R-CNN (Ren et al. 2015). For ViPer, we further preserve the fine-grained vision and language features $\mathbf{Z}_t \in \mathbb{R}^{N\times C}$ and $\mathbf{Z}_i \in \mathbb{R}^{HW\times C}$ for matching the fine-grained vision and language features under visual perturbations.

Spatial ViPer

A key difference between TBPS and text-based person ReID is that the former dynamically constructs the gallery for retrieval from the detection results. Thus the visual features can be obtained from misaligned person boundaries while the language features are given by well-aligned descriptions. For this, we propose Spatial ViPer that perturbs the region proposals to obtain visual features with misaligned boundaries and learn to match the well-aligned language features

with the visual features.

Specifically, we perform center shifting and box scaling (Li et al. 2022a) on each GT bounding box to obtain n perturbed region proposals. Denoted by (c_x, c_y) and (h, w) the center point and spatial size of a GT box, center shifting adds a random shift $(\Delta c_x, \Delta c_y)$ to the center, where $|\Delta c_x| < \frac{\lambda_1 w}{2}$ and $|\Delta c_y| < \frac{\lambda_1 h}{2}$. And box scaling randomly samples height and width from $[(1-\lambda_2)\,h, (1+\lambda_2)\,h]$ and $[(1-\lambda_2)\,w, (1+\lambda_2)\,w]$. $\lambda_1 \in (0,1)$ and $\lambda_2 \in (0,1)$ are hyper-parameters that control the scale of Spatial ViPer.

Although Spatial ViPer is designed only for learning person retrieval features, matching vision and language features implicitly unifies visual features from differently perturbed proposals and blurs the boundary information for detection. We thus combine the perturbed proposal with the RPN (Ren et al. 2015) generated region proposals (as in Figure 2b) for detection training.

Attentive ViPer

Spatial ViPer only varies the boundary of instance visual features by perturbed proposals. To further enable visual perturbation within the proposals to simulate interferences such as occlusions and background clutters, a straightforward solution is to randomly drop (Luo et al. 2019) or mix (Zhang et al. 2018) partial of the visual clues before producing the output features. Yet dropping the unattended background clues is unfavourable and mixing discriminative information causes ambiguous matches with the language descriptions. We thus propose to perform attentive token removal and exchange, *i.e.* Attentive ViPer, on the instance visual feature maps $\mathsf{F}^k \in \mathbb{R}^{D\times H\times W}$ as in Figure 2b.

Typically, the highly attended tokens are corresponded with discriminative foreground clues and the less attended tokens mostly indicate background regions. To estimate the model attention on the visual tokens without extra annotation, we employ the self-attention layer in the image projector. Formally, the self-attention image projector combines the flattened instance feature map $\mathbf{F}^k \in \mathbb{R}^{HW\times D}$ and the mean vector $\mathbf{f}^k \in \mathbb{R}^D$ of $\mathbf{F}^k$ to form an input sequence $[\mathbf{f}^k, \mathbf{F}^k]$, where $\mathbf{f}^k$ serves as the [CLS] token (Dosovitskiy et al. 2020). Thus the attention $\mathbf{a}^k \in \mathbb{R}^{HW}$ between $\mathbf{f}^k$ and $\mathbf{F}^k$ calculated within the self-attention layer reflects model attention on the tokens to produce the global visual feature. Note that A^k is only used to guide Attentive ViPer and doesn't backpropagate gradients.

Given $\mathbf{a}^k$, for highly attended tokens, we select the top-n_r attended tokens, where n_r is randomly sampled from $[m_r^0, m_r^1]$ and remove them within the image projector when producing the global visual features. For less attended tokens, to avoid mixing discriminative person features, we take the L2 distance between attentions as the match cost and employ Hungarian matching between instances in a batch to exchange the bottom-n_e attended tokens, where n_e is randomly sampled from $[m_e^0, m_e^1]$. Given the following definitions:

- $\mathsf{GlobalProj}(\mathbf{F}^k, \mathbf{m})$: The module extracts the global feature of $\mathbf{F}^k$ with a binary vector $\mathbf{m} \in \{0,1\}^{HW}$ indicates whether to ignore each visual token in self-attention;
- $\mathsf{argmax}(\mathbf{a}, n)$ and $\mathsf{argmin}(\mathbf{a}, n)$: The indices of the top-n and bottom-n elements in vector $\mathbf{a}$, respectively;
- $\mathsf{MinL2Match}(\mathcal{A})$: The indices of the optimally matched peers of the elements in $\mathcal{A}$ with L2 matching distance;

the detailed algorithm of Attentive ViPer is described in Algorithm 1.

Algorithm 1: Attentive ViPer

Input: Instance features $\mathcal{F} = \{\mathbf{F}^k | k = 1, \ldots, K\}$
Parameter: min/max removable tokens m_r^0/m_r^1, min/max exchangeable tokens m_e^0/m_e^1
Output: Global features $\mathcal{G}$

Let $\mathcal{G} = \emptyset$.
/* token unchanged */
for $\mathbf{F}^k$ in $\mathcal{F}$ **do**
  $\mathcal{G} = \mathcal{G} \cup \{\mathsf{GlobalProj}(\mathbf{F}^k, \mathbf{1})\}$
end for
/* token remove */
calculate attentions $\mathcal{A}_r$ on $\mathcal{F}$
for $\mathbf{F}^k$ in $\mathcal{F}$ and $\mathbf{a}^k$ in $\mathcal{A}_r$ **do**
  $\mathbf{m} = \mathbf{1}, |\mathbf{m}| = HW$
  randomly sample $n_r \in [m_r^0, m_r^1]$
  $\mathbf{m}[\mathsf{argmax}(\mathbf{a}^k, n_r)] = 0$,
  $\mathcal{G} = \mathcal{G} \cup \{\mathsf{GlobalProj}(\mathbf{F}^k, \mathbf{m})\}$
end for
/* token exchange */
calculate attentions $\mathcal{A}_e$ on $\mathcal{F}$
$B = \mathsf{MinL2Match}(\mathcal{A}_e), |B| = |\mathcal{F}|$
for $\mathbf{F}^k$ in $\mathcal{F}$, $\mathbf{a}^k$ in $\mathcal{A}_e$ and b in B **do**
  randomly sample $n_e \in [m_e^0, m_e^1]$
  $\mathbf{F}^k[\mathsf{argmin}(\mathbf{a}^k, n_e)] = \mathbf{F}^b[\mathsf{argmin}(\mathbf{a}^k, n_e)]$
  $\mathcal{G} = \mathcal{G} \cup \{\mathsf{GlobalProj}(\mathbf{F}^k, \mathbf{1})\}$
end for
return $\mathcal{G}$.

Fine-grained ViPer

As fine-grained vision-language alignment (Zhang et al. 2023; Su et al. 2024; Jiang and Ye 2023) is typically incorporated for text-based person retrieval, we further introduce visual perturbations at the fine granularity. While the projected language features $\mathcal{Z}_t \in \mathbb{R}^{N\times C}$ are kept clear, we randomly replace a ratio r of the visual tokens in the paired $\mathbf{F}^k$ with a learned [MASK] token and send the masked visual feature sequence to the image projector to produce the projected fine-grained sequence $\mathcal{Z}_i \in \mathbb{R}^{HW\times C}$. We take $\mathcal{Z}_i$ as the query and the paired language features as the key/value to a cross-modal transformer (Vaswani et al. 2017; Dosovitskiy et al. 2020) as in Figure 2c. Different from masked image modeling (He et al. 2022; Xie et al. 2022), the transformer gradually restores the masked visual features from the correlated language features, facilitating fine-grained vision-language alignment under visual perturbations.

We also note that IRRA (Jiang and Ye 2023) designed an MLM-based mechanism that recovers masked words by stacked sefl-attention layers on fused multi-modal features. Yet this still allows reconstruction of masked information

within the masked modality (Devlin et al. 2018; Xie et al. 2022), we thus employ a full cross-attention architecture to suppress inner-modality masked visual information modeling and enforces the masked fine-grained visual features to be recovered from the paired language features. Formally, the full cross-attention transformer consists of M cross-attention layers defined by

$$\begin{aligned}\mathbf{H}'_m &= \mathrm{MHA}\left(\mathrm{LN}\left(\mathbf{H}_{m-1}\right), \mathbf{Z}_t, \mathbf{Z}_t\right) + \mathbf{H}_{m-1} \\ \mathbf{H}_m &= \mathrm{FFN}(\mathrm{LN}\left(\mathbf{H}'_m\right)) + \mathbf{H}'_m\end{aligned} \tag{1}$$

where $m \in \{1, 2, \ldots, M\}$ and $\mathbf{H}_0 = \mathbf{Z}_i$. MHA refers to the multi-headed attention (Vaswani et al. 2017; Dosovitskiy et al. 2020) and LN denotes the layer normalization. FFN is a two-layer feedforward network with an intermediate GELU activation function. The final output is $\tilde{\mathbf{F}}^k = \mathrm{Linear}(\mathbf{H}_M)$, $\tilde{\mathbf{F}}^k \in \mathbb{R}^{HW \times D}$. The overall Fine-grained ViPer is supervised by an MSE loss

$$\mathcal{L}_{mse} = \frac{1}{HW} ||\tilde{\mathbf{F}}^k - \mathbf{F}^k||_2 \tag{2}$$

to predict the masked visual tokens. Due to the overlapped proposals in the Faster R-CNN pipeline, the number of duplicated visual feature sequences of a person is likely to exceed the number of language feature sequences. Thus for each language feature sequence, we randomly select one matched visual feature sequence for training.

Training and Inference

During training, Spatial ViPer has an impact on both Attentive ViPer and Fine-grained ViPer as both the global and fine-grained visual features are obtained based on the proposals. Attentive ViPer and Fine-grained ViPer are conducted in parallel for the stability of training, *i.e.* the unperturbed fine-grained visual features are optimized by Fine-grained ViPer. The overall training objective is defined by

$$\mathcal{L} = \mathcal{L}_{det} + \mathcal{L}_{oim} + \mathcal{L}_{cmpm} + \mathcal{L}_{mse}. \tag{3}$$

For inference, we only extract global vision and language person features for person retrieval. The cosine distance between cross-modal features is calculated to measure the similarities between language queries and detected persons in the gallery images.

Experiments

Implementation Details

In the proposed model, the hyperparameters for detection are set following previous works (Chen et al. 2020b; Li and Miao 2021) except that the output spatial size of ROI align is 16×8. The cross-modal transformer contains 4 multi-head cross-attention layers with the number of heads set to 16. For training, we set the batch size to be 8 and employ a multi-scale training strategy as in previous IBPS models (Chen et al. 2020b; Yan et al. 2021). The model is optimized with the Adam optimizer and an initial learning rate of $1e$-5 which is linearly warmed up during the first two epochs. We train the model for 20 epochs and decrease the learning rate by 10 at the 12-th epoch for CUHK-SYSU-TBPS. For PRW-TBPS, the model is trained for 25 epochs and the learning rate is decayed by 10 at the 12-th epoch. In the OIM loss, the circular queue sizes are 5000 for CUHK-SYSU-TBPS and 500 for PRW-TBPS. The temperature σ for the OIM losses is set to $1/30$ and the momentum coefficient is 0.5 following (Chen et al. 2020b; Li and Miao 2021). λ_1, λ_2 and n for Spatila Viper are set to 0.4, 0.2 and 3, respectively. The min/max removable and exchangeable tokens m_r^0/m_r^1 and m_e^0/m_e^1 are 1/4 and 4/8, respectively. The masking ratio r in Fine-grained ViPer is 0.5. At test time, we rescale the test images to a fixed size of 1500×900 pixels. All experiments are conducted on a single RTX 3090 GPU.

Method	PRW-TBPS		CUHK-SYSU-TBPS	
	mAP	top-1	mAP	top-1
OIM (Xiao et al. 2017)+BiLSTM	4.58	6.66	23.74	17.42
NAE (Chen et al. 2020b)+BiLSTM	5.20	7.54	23.48	16.62
BSL (Zhang et al. 2023) +BiLSTM	3.60	6.42	26.91	20.97
OIM (Xiao et al. 2017)+BERT	8.52	14.44	43.39	36.59
NAE (Chen et al. 2020b)+BERT	9.20	14.44	45.70	39.14
BSL (Zhang et al. 2023)+BERT	10.70	16.82	48.39	40.83
SDPG (Zhang et al. 2023)	11.93	21.63	50.36	49.34
MACA (Su et al. 2024)	18.18	33.25	57.77	52.03
ViPer (ours)	**22.07**	**35.62**	**62.13**	**55.82**

Table 1: Performance comparisons between previous TBPS methods and our proposed model. The models are evaluated on the PRW-TBPS and CUHK-SYSU-TBPS datasets. All compared TBPS results are drawn from SDPG (Zhang et al. 2023) and MACA (Su et al. 2024).

Datasets

PRW-TBPS is collected based on the IBPS dataset PRW (Zheng et al. 2017). Each box of the labeled persons in the training and query set is annotated with one or more sentences. The training set contains 5,704 scene images captured by multiple cameras deployed on campus. A total of 483 different person identities and 14,897 boxes are densely labeled. The query set presents independently annotated sentences matched with 2,057 query person boxes of 450 different identities. For evaluation, a total of 6,112 scene images without overlapping with the training set is employed for query person retrieval.

CUHK-SYSU-TBPS is based on the IBPS dataset CUHK-SYSU (Xiao et al. 2017) and the text-based perosn ReID dataset CUHK-PEDES (Li et al. 2017). 11,206 training scene images with 15,080 person boxes of 5,532 persons and 6,978 gallery images of 2,900 persons are presented. The language annotations in CUHK-PEDES are reused for the labeled persons in the training set and query set, which equips each box with two independent sentences. Different from PRW-TBPS, CUHK-SYSU-TBPS follows CUHK-SYSU to present various test settings with gallery sizes varying from 50 to 4000 to examine the model capability.

Evaluation protocols of person search share a similar spirit of that in person re-identification (Luo et al. 2019; Ye et al. 2021; Jiang and Ye 2023). The widely adopted mAP and

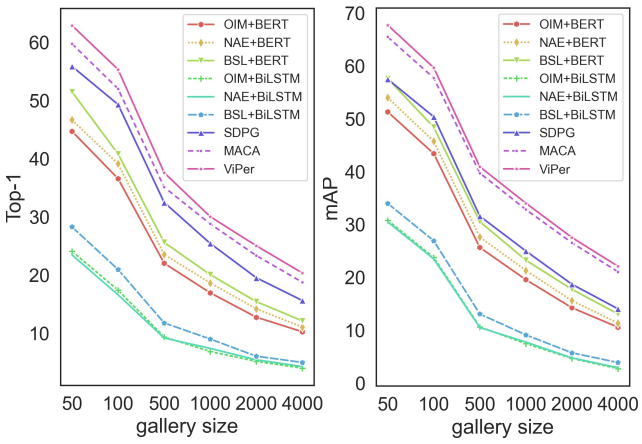


Figure 3: Person search performances on CUHK-SYSU-TBPS under various gallery sizes.

top-1 accuracy are utilized as the evaluation metrics for person search. During evaluation, a bounding box in a gallery image is considered a true positive if it shares the same identity label with the query and has an intersection-over-union (IOU) larger than 0.5 with the ground truth bounding box. In this way, the actual number of gallery person boxes is determined by the person detection result. Both the mAP and top-1 accuracy of the person search results will be affected by the person detection performance.

Comparison with SOTA

PRW-TBPS As is shown in Table 1, the proposed method achieves the best mAP score of 22.07% and the top-1 accuracy of 35.62% on the PRW-TBPS dataset among TBPS methods. Compared with the second best MACA (Su et al. 2024), the results are clearly improved by 3.89% mAP and 2.37% top-1. It is also worth noting that compared with recent IBPS methods (Li and Miao 2021; Lee et al. 2022; Cao et al. 2022; Yu et al. 2022), the performances of TBPS models are still largely inferior, suggesting that there is still large room to improve the TBPS accuracy.

CUHK-SYSU-TBPS On the CUHK-SYSU-TBPS dataset, the proposed model consistently surpasses previous TBPS methods, especially on the top-1 score, as presented in Table 1. Similar to that on PRW, though the smaller gallery size of CUHK-SYSU-TBPS makes a simpler person search task than PRW-TBPS, the results of TBPS methods still significantly fall behind the IBPS counterparts. To better understand the effectiveness of the proposed model, we additionally test the model performances under varied gallery sizes as in Figure. 3. The results of the compared TBPS methods are drawn from MACA (Su et al. 2024). It can be observed that the proposed model consistently outperforms previous models by a clear margin as the gallery size grows, demonstrating the robustness of the proposed method.

Ablation Study

To understand the effect of each proposed module, we perform ablation studies on the PRW-TBPS dataset by sequentially adding the modules on the baseline model. The main evaluation results are collected in Table 2. More analytical experiments are presented in the Supplementary Material.

Spatial ViPer	Attentive ViPer		Fine-grained ViPer		mAP	top-1
	token remove	token exchange	IRRA Encoder	proposed		
					19.41	31.25
✓					20.26	32.26
✓	✓				20.75	34.10
✓		✓			20.82	34.08
✓	✓	✓			21.42	34.94
✓	✓	✓	✓		21.75	34.71
✓	✓	✓		✓	22.07	35.62

Table 2: Effect of each component of the proposed method.

Token Remove		Token Exchange		mAP	top-1
attentive	random	attentive	random		
				20.26	32.26
✓				20.75	34.10
	✓			20.15	33.53
		✓		20.82	34.08
			✓	20.59	32.77

Table 3: Comparison between attentive and random token manipulation.

The effect of Spatial ViPer. As presented in the first two rows of Table 2, we compare the baseline model with and without the proposed Spatial Viper. It can be observed that integrating Spatial Viper in the baseline model clearly boosts the TBPS performance, demonstrating its effectiveness. Compared with MACA (Su et al. 2024), although the overall architecture of the proposed baseline model is simpler, we observe that the baseline already achieves superior TBPS performance. This mainly benefits from the pre-trained vison-language representation in CLIP (Radford et al. 2021).

The effect of Attentive ViPer. To verify the effectiveness of Attentive ViPer, we first test to separately employ the proposed token remove and exchange as in the 3rd and 4th rows of Table 2. It can be observed that both modules improve the performance compared with Spatial Viper. We then test to integrate the overall Attentive ViPer on top of Spatial Viper. This consistently enhances the TBPS performance as in the 5th row of Table 2. As Attentive ViPer relies on attention to perform token manipulation, we also test random manipulations in Table 3 to understand the effect of attention guidance. Compared with the model without token manipulations, adding any of the random manipulations only slightly boosts the TBPS results, while the proposed attentive modules show clear improvements, suggesting that the attention guidance also plays a vital role in Attentive ViPer.

The effect of Fine-grained ViPer. On top of Spatial ViPer and Attentive ViPer, we further integrate Fine-grained ViPer as in the last two rows of Table 2. The IRRA Encoder (Jiang and Ye 2023) is also tested as the cross-modal transformer for comparisons. It can be observed that Fine-grained ViPer consistently improves the TBPS performance com-

Cross-modal Transformer	Param.(M)	RSTPReid		
		mAP	top-1	mINP
IRRA Encoder	54.58	37.81	48.75	17.02
proposed	**50.38**	**38.45**	**49.95**	**17.55**

Table 4: Performance comparisons between the Encoder in IRRA (Jiang and Ye 2023) and our proposed cross-modal transformer on the text-based person ReID dataset RSTPReid (Zhu et al. 2021).

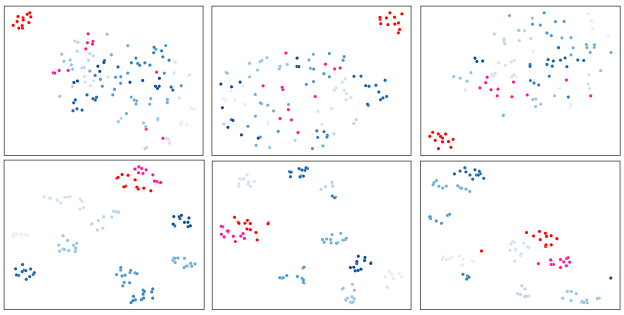


Figure 4: Examplar illustration of vision and language feature distribution of the proposed model. We compare our trained model with CLIP (Radford et al. 2021) pre-trained parameters.

pared with the models supervised by only global feature learning. When comparing the full cross-attention transformer with IRRA Encoder, the model gains further enhancement for TBPS.

Moreover, we test to take the full cross-attention transformer as an alternative to IRRA Encoder for text-based person ReID. As shown in Table 4, the experiments are conducted on a popular text-based person ReID dataset RSTPReid (Zhu et al. 2021). For comparison, we employ the overall IRRA model. The CLIP (Radford et al. 2021) pre-trained ResNet50 (He et al. 2016) is used to initialize the image encoder. Compared with IRRA Encoder, our proposed cross-modal transformer reduces the number of parameters while improving the person ReID performance, which further demonstrates the effect of suppressing inner-modality masked information reconstruction by the full cross-attention architecture.

Visualization

Visualization of feature distribution. To qualitatively understand the distribution of the vision and language person features, we present the t-SNE visualization of three example persons as in Figure 4. Specifically, we randomly select 10 identities from PRW-TBPS and 1 identity among them as the anchor. In Figure 4, for instances of the anchor identity, we use red points to indicate the language features of the instances and magenta points to represent their visual features. For instances of the rest identities, we denote by gradient blue colors their visual features. For illustration, we compare the feature distributions of the same instances from CLIP (the 1st row) and ViPer (the 2nd row). It can be observed that the features drawn from CLIP pre-trained models tend to be ambiguous for TBPS. In contrast, ViPer effectively narrows the distances between cross-modal positive pairs and pushes away the distractors.

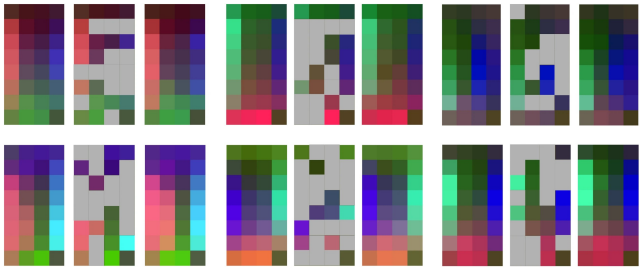


Figure 5: Visualization of visual tokens of 6 random persons in Fine-grained Viper. We group the **original**, **masked**, and **reconstructed** tokens of the same person as a triplet and conduct PCA to obtain their 3-channel representations. The masked tokens are marked with gray-colored squares.

Visualization of restored tokens in Fine-grained ViPer. We also conduct visualization of the person visual tokens in Fine-grained ViPer as in Figure 5. Concretely, we group the original, masked, and reconstructed visual tokens of the same person and employ PCA to reduce the dimensionality of the features for visualization. As Figure 5 shows, the original feature maps are randomly masked with a learned token. By fusing the multi-modal information through cross-attention layers, the reconstructed visual tokens can be observed to be well-matched with their unmasked version.

Conclusion

This work proposes a Visual Perturbation network for TBPS to tackle the discrepency between training and inference. We design three visual perturbation modules based on a CLIP-driven baseline network to force vision-language alignment under the condition that the visual features encode only partial language-described clues of the same person. Specifically, Spatial ViPer is proposed to produce visual features with misaligned boundaries by varying proposals. On top of that, we introduce Attentive ViPer to adaptively manipulate tokens based on attention before aggregating them, enabling visual perturbations on global visual features. For fine-grained visual features, we design Fine-grained ViPer that randomly masks and recovers a ratio of visual tokens from correlated language clues. A full cross-attention transformer is employed to perform cross-modal interactions without inner-modality masked information reconstruction. Experimental results show that the proposed model outperforms previous methods on existing TBPS datasets. Analytical experiments are conducted to verify the effectiveness of the proposed modules. Through the results, we also observe that the performances of TBPS models are still marginally inferior compared with their IBPS counterparts. Future works for TBPS should explore to further enhance the discriminative vision-language feature learning to boost the performance.

Acknowledgments

This work is supported by the National Natural Science Foundation of China 62276016 and 62372029.

References

Cao, J.; Pang, Y.; Anwer, R. M.; Cholakkal, H.; Xie, J.; Shah, M.; and Khan, F. S. 2022. PSTR: End-to-End One-Step Person Search With Transformers. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*.

Cao, M.; Bai, Y.; Zeng, Z.; Ye, M.; and Zhang, M. 2024. An empirical study of clip for text-based person search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 465–473.

Chen, D.; Zhang, S.; Ouyang, W.; Yang, J.; and Schiele, B. 2020a. Hierarchical Online Instance Matching for Person Search. In *AAAI*.

Chen, D.; Zhang, S.; Yang, J.; and Schiele, B. 2020b. Norm-aware embedding for efficient person search. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12615–12624.

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

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

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

Han, C.; Zheng, Z.; Gao, C.; Sang, N.; and Yang, Y. 2021. Decoupled and memory-reinforced networks: Towards effective feature learning for one-step person search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 1505–1512.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16000–16009.

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

Jiang, D.; and Ye, M. 2023. Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2787–2797.

Lan, X.; Zhu, X.; and Gong, S. 2018. Person search by multi-scale matching. In *Proceedings of the European conference on computer vision (ECCV)*, 536–552.

Lee, S.; Oh, Y.; Baek, D.; Lee, J.; and Ham, B. 2022. OIM-Net++: Prototypical Normalization and Localization-aware Learning for Person Search. In *European Conference on Computer Vision*. Springer.

Li, F.; Zhang, H.; Liu, S.; Guo, J.; Ni, L. M.; and Zhang, L. 2022a. Dn-detr: Accelerate detr training by introducing query denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13619–13627.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022b. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, 12888–12900. PMLR.

Li, L. H.; Zhang, P.; Zhang, H.; Yang, J.; Li, C.; Zhong, Y.; Wang, L.; Yuan, L.; Zhang, L.; Hwang, J.-N.; et al. 2022c. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10965–10975.

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

Li, Z.; and Miao, D. 2021. Sequential end-to-end network for efficient person search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 2011–2019.

Luo, H.; Gu, Y.; Liao, X.; Lai, S.; and Jiang, W. 2019. Bag of tricks and a strong baseline for deep person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 0–0.

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

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28: 91–99.

Su, L.; Quan, R.; Qi, Z.; and Qin, J. 2024. MACA: Memory-aided Coarse-to-fine Alignment for Text-based Person Search. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2497–2501.

Suo, W.; Sun, M.; Niu, K.; Gao, Y.; Wang, P.; Zhang, Y.; and Wu, Q. 2022. A simple and robust correlation filtering method for text-based person search. In *European conference on computer vision*, 726–742. Springer.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*.

Wang, C.; Ma, B.; Chang, H.; Shan, S.; and Chen, X. 2020. Tcts: A task-consistent two-stage framework for person search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11952–11961.

Xiao, T.; Li, S.; Wang, B.; Lin, L.; and Wang, X. 2017. Joint detection and identification feature learning for person search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3415–3424.

Xie, Z.; Zhang, Z.; Cao, Y.; Lin, Y.; Bao, J.; Yao, Z.; Dai, Q.; and Hu, H. 2022. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9653–9663.

Yan, S.; Dong, N.; Zhang, L.; and Tang, J. 2023. Clip-driven fine-grained text-image person re-identification. *IEEE Transactions on Image Processing*.

Yan, Y.; Li, J.; Qin, J.; Bai, S.; Liao, S.; Liu, L.; Zhu, F.; and Shao, L. 2021. Anchor-free person search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7690–7699.

Yang, F.; Li, W.; Yang, M.; Liang, B.; and Zhang, J. 2024. Multi-Modal Disordered Representation Learning Network for Description-Based Person Search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 16316–16324.

Ye, M.; Shen, J.; Lin, G.; Xiang, T.; Shao, L.; and Hoi, S. C. 2021. Deep learning for person re-identification: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6): 2872–2893.

Yu, R.; Du, D.; LaLonde, R.; Davila, D.; Funk, C.; Hoogs, A.; and Clipp, B. 2022. Cascade Transformers for End-to-End Person Search. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Zhang, H.; Cisse, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2018. mixup: Beyond Empirical Risk Minimization. In *International Conference on Learning Representations*.

Zhang, S.; Cheng, D.; Luo, W.; Xing, Y.; Long, D.; Li, H.; Niu, K.; Liang, G.; and Zhang, Y. 2023. Text-based person search in full images via semantic-driven proposal generation. In *Proceedings of the 4th International Workshop on Human-centric Multimedia Analysis*, 5–14.

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

Zheng, L.; Zhang, H.; Sun, S.; Chandraker, M.; Yang, Y.; and Tian, Q. 2017. Person re-identification in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1367–1376.

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

Zuo, J.; Zhou, H.; Nie, Y.; Zhang, F.; Guo, T.; Sang, N.; Wang, Y.; and Gao, C. 2024. UFineBench: Towards Text-based Person Retrieval with Ultra-fine Granularity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22010–22019.