An Empirical Study of CLIP for Text-Based Person Search

Min Cao[1], **Yang Bai** [1], **Ziyin Zeng** [1], **Mang Ye** [2*], **Min Zhang** [3]

[1] School of Computer Science and Technology, Soochow University
[2] School of Computer Science, Wuhan University
[3] Harbin Institute of Technology, Shenzhen
{caomin0719@126.com, mangye16@gmail.com}

Abstract

Text-based Person Search (TBPS) aims to retrieve the person images using natural language descriptions. Recently, Contrastive Language Image Pretraining (CLIP), a universal large cross-modal vision-language pre-training model, has remarkably performed over various cross-modal downstream tasks due to its powerful cross-modal semantic learning capacity. TPBS, as a fine-grained cross-modal retrieval task, is also facing the rise of research on the CLIP-based TBPS. In order to explore the potential of the visual-language pre-training model for downstream TBPS tasks, this paper makes the first attempt to conduct a comprehensive empirical study of CLIP for TBPS and thus contribute a straightforward, incremental, yet strong TBPS-CLIP baseline to the TBPS community. We revisit critical design considerations under CLIP, including data augmentation and loss function. The model, with the aforementioned designs and practical training tricks, can attain satisfactory performance **without any sophisticated modules**. Also, we conduct the probing experiments of TBPS-CLIP in model generalization and model compression, demonstrating the effectiveness of TBPS-CLIP from various aspects. This work is expected to provide empirical insights and highlight future CLIP-based TBPS research.The code is available at https://github.com/Flame-Chasers/TBPS-CLIP.

1 Introduction

Text-based Person Search (TBPS) retrieves the person images from a large-scale image database given a textual description. It is gradually gaining extensive attention (Jiang and Ye 2023; Bai et al. 2023a) due to its potential applications in searching for suspects, locating lost children, *etc*. As a fine-grained retrieval task, it shows challenges in achieving effective cross-modal alignment and efficient cross-modal retrieval, which are crucial for practical applications.

In reply to these challenges, many methods (Zhang and Lu 2018; Ding et al. 2021) focus on projecting representations extracted from each modality into one shared space. However, most of them only utilize unimodal pre-trained models as the backbones and ignore the powerful Vision-Language Pre-training (VLP) models equipped with an adequate understanding of cross-modal alignment. Recently, VLP methods (Lu et al. 2019; Li et al. 2022) have verified excellent performance on various cross-modal downstream tasks. Consequently, it has become the dominant paradigm for solving these tasks. Considering that the research under VLP for TBPS (Bai et al. 2023b,a) is now in its infancy, this paper aims to fully exploit the potential of VLP for TBPS.

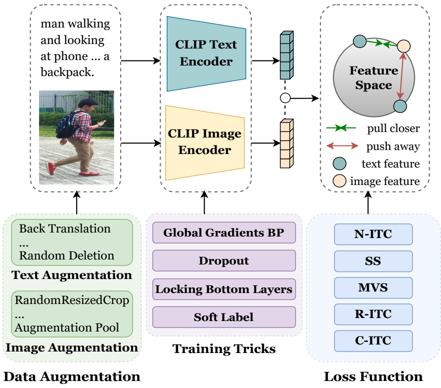


Figure 1: Overview of the empirical study on CLIP.

CLIP (Radford et al. 2021) stands out among VLP methods and has verified impressive performance on various downstream tasks (Wang et al. 2022b,a). Notably, CLIP offers efficient retrieval capabilities by encoding each modality independently. These inspire some researchers to explore CLIP-based TBPS methods (Wang et al. 2023; Jiang and Ye 2023). These methods focus on combining complex modules into CLIP to improve performance, lacking the full exploitation of CLIP's pre-trained knowledge and powerful cross-modal semantic learning capacity. This paper refocuses on the essence of CLIP and explores its fine-tuning potential for TBPS. We conduct a thorough empirical study of CLIP for TBPS from two perspectives:

1) Data augmentation. It is a prevalent and effective technique to increase the generalization of models by learning augmentation-invariant representations between the original input and the augmented version. Until now, none of the TBPS methods have deeply and systematically explored this technique. Instead, they typically employ a simple utilization of random horizontal flip for the image while not applying any augmentation to the text (Wang et al. 2020; Chen et al. 2022; Ding et al. 2021). In this study, we comprehen-

[*] Corresponding author

sively evaluate various data augmentation strategies, thereby deriving a powerful data augmentation strategy for TBPS.

2) Loss function. Designing rational and practical loss functions is critical to improving performance and has been an increasingly active research direction in TBPS community (Zhang and Lu 2018; Bai et al. 2023a). We take CLIP as a hotbed and conduct a series of probing studies to analyze the effectiveness of various loss functions in TBPS. Unlike the loss functions in existing TBPS methods that are well-designed mainly from exploring the TBPS task and belong to the task-oriented loss functions, the loss functions probed in this study are primarily inspired by VLP communities and are pretty generic to various cross-modal tasks.

These empirical studies above, combined with other valuable tricks detailed in later sections, enable us to develop a strong TBPS-CLIP baseline. Unlike other methods of designing sophisticated modules, TBPS-CLIP attains competitive performance under a very lightweight and low-cost architecture. It blends only a few common training tricks, data augmentations, and loss functions into CLIP. To thoroughly verify the effectiveness and generalization of TBPS-CLIP, we further develop some valuable probing experiments.

1) Model generalization. For one thing, CLIP has become a baseline of various tasks due to its effectiveness and simplicity, likewise, TBPS-CLIP is targeted as the baseline of TBPS, for which its effectiveness as the baseline is experimentally proved. For another thing, beyond only employing CLIP for the supervised TBPS, we provide a preliminary exploration of the few-shot TBPS under CLIP and the superiority of TBPS-CLIP in the few-shot setting is also proved.

2) Model compression. We provide insights into the internal properties of TBPS-CLIP by investigating the contribution of each module to the final retrieval performance. These insights offer the guidance for compression of TBPS-CLIP.

In short, this paper contributes a strong TBPS-CLIP baseline to the TBPS community. TBPS-CLIP offers high-performance, lightweight, cost-effective architecture and ease of use. We also show its advantage in model generalization and model compression. All these manifest the applicability of TBPS-CLIP in both academia and industry.

2 Related Work

2.1 Text-based Person Search

TBPS is closely connected with person re-identification (Ye et al. 2021) and image-text retrieval (Cao et al. 2022). Conventional TBPS methods mainly adopt unimodal pre-trained models as backbones (Wang et al. 2020; Wu et al. 2021; Li et al. 2017a; Li, Cao, and Zhang 2022). These methods typically use ResNet-50 or ViT as the image encoder and LSTM or BERT as the text encoder. Some works (Zhang and Lu 2018; Sarafianos, Xu, and Kakadiaris 2019; Li et al. 2017a; Li, Cao, and Zhang 2022) design specific losses to learn discriminative representations based on the unimodal backbones. Some extract fine-grained information from images and text to align cross-modal fine-grained features. For instance, partitioning images into horizontal stripes is a technique used to obtain fine-grained image information (Chen et al. 2022; Zheng et al. 2020; Ding et al. 2021), while fine-grained text information can be obtained by parsing a set of noun phrases using the NLTK Toolbox (Wang et al. 2020).

Witnessing VLP's great success on cross-modal tasks in recent years, researchers (Han et al. 2021; Yan et al. 2022; Wang et al. 2023; Jiang and Ye 2023; Bai et al. 2023a) have begun pushing the frontier of TBPS solutions with VLP. Han *et al.* (Han et al. 2021) enhance text feature encoding by using CLIP (Radford et al. 2021) as the backbone and merging a Bi-GRU after its text encoder; CFine (Yan et al. 2022) leverages CLIP image encoder to enhance cross-modal correspondence, replacing the text encoder with BERT to prevent distortion of intra-modal information; TP-TPS (Wang et al. 2023) probes textual capabilities of CLIP by aligning images with multi-integrity descriptions and attribute prompts; IRRA (Jiang and Ye 2023) designs an implicit relation reasoning module above CLIP to align fine-grained information across modalities; RaSa (Bai et al. 2023a) develops two novel losses under ALBEF (Li et al. 2021a); Bai *et al.* (Bai et al. 2023b) leverage VLP knowledge to solve TBPS without parallel image-text data. Unlike these methods that exploit specific modules beyond VLP to improve performance, this paper aims to maximize the potential of CLIP itself for TBPS without relying on additional modules, resulting in TBPS-CLIP with competitive performance.

2.2 Vision-Language Pre-training

In recent years, VLP (Lu et al. 2019; Li et al. 2022) has emerged as a dominant solution for various cross-modal tasks. VLP methods leverage large-scale pre-training on abundant image-text pairs, enabling the models to learn robust cross-modal representations that can be further fine-tuned for different downstream tasks. Among various VLP methods, CLIP (Radford et al. 2021) stands out for its excellent performance (Wang et al. 2022b,a). Furthermore, several works are carried out to enhance the data efficiency of CLIP. For example, SLIP (Mu et al. 2022) introduces a self-supervised learning loss alongside the contrastive loss in CLIP; FILIP (Yao et al. 2021) employs a token-wise maximum similarity between visual and textual tokens to guide the contrastive loss in CLIP; DeCLIP (Li et al. 2021b) exploits multiple supervisions, including self-supervision, multi-view supervision and nearest-neighbor supervision, to replace the single contrastive supervision. In addition, some works concentrate on exploring CLIP's few-shot capabilities. For example, CoOp (Zhou et al. 2022) utilizes learnable vectors as the prompts and achieves significant performance gain in a few-shot regime, and CLIP-Adapter (Gao et al. 2021) appends the learnable module to CLIP and achieves decent few-shot results by only fine-tuning this module.

3 Empirical Studies

Grounded on CLIP, we first introduce some practical training tricks to strengthen the CLIP baseline in Section 3.1, and then elaborate data augmentations and loss functions for the empirical study in Section 3.2 and Section 3.3, respectively. Model generalization is examined in Section 3.4, and model compression is discussed in Section 3.5 *Additional details on the following contents can be found in the Appendix.* The overview of the model is illustrated in Figure 1.

3.1 Training Tricks

We investigate four common training tricks around CLIP. They are: *global gradients back-propagation*, *dropout*, *locking bottom layers* and *soft label*.

3.2 Data Augmentations

Image Augmentation. We classify image augmentations into two groups: removal and alteration. The first group has operations that remove some information from the image, including *RandomResizedCrop*, *RandomErasing*, *RandomGrayscale* and *GaussianBlur*. The second alters the color or orientation of the image with keeping the main content, including *ColorJitter*, *RandomHorizontalFlip*, *RandomVerticalFlip* and *RandomRotation*.

Using multiple augmentations simultaneously can introduce significant distortion to the original image, leading to a decline in performance. Given that, we explore another set of augmentations. (1) *AutoAugment (Cubuk et al. 2018)* automatically searches for the best augmentation policy using reinforcement learning. We use its default settings on PyTorch in the experiment. (2) *RandAugment (Cubuk et al. 2020)* removes the searching stage in AutoAugment and instead randomly selects from a pool of augmentation operations. It involves two hyperparameters: the number of augmentations, denoted as N, and the magnitude, denoted as M. The default settings are $N = 2$ and $M = 9$. (3) *TrivialAugment (Müller and Hutter 2021)* further drops the requirement of setting hyperparameters in RandAugment, and instead randomly selects one augmentation and its magnitude to apply on each image. It is completely parameter-free. (4) *An augmentation pool strategy* is designed in this paper, inspired by the above-mentioned automatic data augmentations. Specifically, two augmentations are randomly applied to each image.

Text Augmentation. Text augmentation options are relatively more limited compared to image augmentation due to the abstract and discrete nature of language. (1) *Back translation* translates the original text to a specific language and then translates it back, by which we can obtain more diverse textual descriptions while preserving its original meaning. We use French as the intermediate language due to its closer linguistic resemblance to English, resulting in fewer semantic changes in the translated back text. (2) *Synonym replacement* randomly selects words from the sentence and replaces with randomly chosen synonyms. (3) *Random insertion* randomly chooses some words from the sentence and the synonyms of selected words are then inserted into random positions of the sentence. (4) *Random swap* selects two words from the sentence at random and interchanges their positions. (5) *Random deletion* randomly removes each word in a sentence. (6) *EDA* (Wei and Zou 2019) randomly selects one from synonym replacement, random insertion, random swap and random deletion and applies it to the sentence.

3.3 Loss Functions

CLIP equips with an image-text contrastive loss to pull positive samples together while pushing negative ones apart. Further, we normalize the label in the loss and obtain the normalized image-text contrastive loss (N-ITC):

$$\mathcal{L}_{N-ITC} = -\frac{1}{2N}(\sum_{i=1}^{N}\sum_{j=1}^{N}\hat{q}_{i,j}\log p_{i,j} + \sum_{i=1}^{N}\sum_{j=1}^{N}\hat{q}_{j,i}\log p_{j,i}), \tag{1}$$

where $\hat{q}_{i,j}$ is normalized by $q_{i,j}/\sum_{k=1}^{N} q_{i,k}$ and $q_{i,j}$ is the ground-truth label (1 for positive pair and 0 for negative one). N is the number of samples, and $p_{i,j}$ represents the pseudo-label that is the probability of matching the image I_i to the text T_j and the reverse applies in $p_{j,i}$,

$$p_{i,j} = \frac{\exp(f_{I_i}\cdot f_{T_j}/\tau)}{\sum_{k=1}^{N}\exp(f_{I_i}\cdot f_{T_k}/\tau)}, p_{j,i} = \frac{\exp(f_{T_i}\cdot f_{I_j}/\tau)}{\sum_{k=1}^{N}\exp(f_{T_i}\cdot f_{I_k}/\tau)}, \tag{2}$$

where f_* is the ℓ_2-normalized representations of the sample, τ is the learnable temperature parameter.

Beyond N-ITC, we study other losses in two directions. One focuses on enhancing data efficiency, and the other targets optimizing the relationship between samples.

Improving Data Efficiency

Self-Supervision. The self-supervised loss (SS) aims to maximize the similarity between two different augmentations of an image and prompts learning robust feature representations with limited data. It has shown effectiveness in various visual tasks (Chen et al. 2020; Chen and He 2021), and motivates us to explore it in TBPS. Specifically,

$$\mathcal{L}_{SS} = -\frac{1}{2N}\sum_{i=1}^{2N}\log\frac{\exp(sim(z_i, z_j)/\tau_s)}{\sum_{k=1,k\neq i}^{2N}\exp(sim(z_i, z_k)/\tau_s)}, \tag{3}$$

where τ_s is a hyper-parameter and set to 0.1, and z_i and z_j are the feature representations of two augmentations of the sample. The SS can be applied on the image or text or both of them, denoted as SS-I, SS-T and SS-IT, respectively.

Multi-View Supervision. The N-ITC in Eq. 1 only leverages one augmented data view. Inspired by DeCLIP (Li et al. 2021b), more views of samples can provide extra supervision to motivate the potential of limited data. Let $\tilde{I}$ and $\tilde{T}$ denote another augmented view of I and T by data augmentation, the N-ITC can be applied on $(\tilde{I}, T)$ or $(I, \tilde{T})$ or $(\tilde{I}, \tilde{T})$, denoted as the multi-view supervision loss of image (MVS-I), multi-view supervision loss of text (MVS-T), multi-view supervision loss of image and text (MVS-IT), respectively.

Optimization for Retrieval

Reversed Image-Text Contrastive Loss. The N-ITC in Eq. 1 approximates optimizing $D_{\mathrm{KL}}(Q\|P)$, with Q as the label distribution and P as the optimized similarity distribution. It mainly focuses on assigning the image-text pair to high similarity probability when its label distribution places a high probability (*i.e.*, positive pair). As a supplement to it, we take inspiration from CMPM (Zhang and Lu 2018) and optimize $D_{\mathrm{KL}}(P\|Q)$, which have P and Q reversed in the N-ITC, and the reversed image-text contrastive loss (R-ITC)

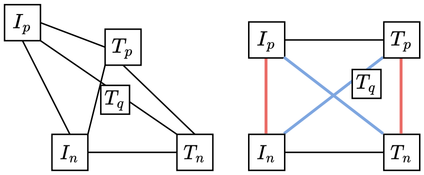


Figure 2: Difference between C-ITC (right-hand) and other matching losses (left-hand), *e.g.*, N-ITC and R-ITC. (I_p, T_p) and (I_n, T_n) are two paired cross-modal data and the query text T_q shares the same identity with (I_p, T_p).

considers separating the negative pairs. Specifically,

$$\begin{aligned}\mathcal{L}_{R-ITC} &= \frac{1}{2}(D_{\mathrm{KL}}(p_{i,j}\|\hat{q}_{i,j}) + D_{\mathrm{KL}}(p_{j,i}\|\hat{q}_{j,i})) \\ &= \frac{1}{2N}(\sum_{i=1}^{N}\sum_{j=1}^{N} p_{i,j}\log\frac{p_{i,j}}{\hat{q}_{i,j}+\epsilon} + \sum_{i=1}^{N}\sum_{j=1}^{N} p_{j,i}\log\frac{p_{j,i}}{\hat{q}_{j,i}+\epsilon}),\end{aligned} \tag{4}$$

where ϵ is a small non-zero value to prevent division by zero.

Cyclic Image-Text Contrastive Loss. The general matching loss (*e.g.*, N-ITC and R-ITC) aims to optimize the relationship between the image and text, which may cause the image and text to be irregularly positioned in the representation space, and carries the risk that the query text T_q mistakenly retrieves the negative image I_n instead of the intended positive image I_p, as illustrated on the left-hand side of Figure 2. Following CyCLIP (Goel et al. 2022), which enhances geometrical consistency in data representation space, we study the cyclic image-text contrastive loss (C-ITC) to mitigate the above problem. The right-hand side of Figure 2 shows the geometry of the resulting representation space.

Specifically, the learned representations are regularized in two ways: in-modality and cross-modality. For the in-modality regularization, the loss enables reducing the gap between the distance of the similarity of two images and that between the corresponding texts:

$$\mathcal{L}_{C^I-ITC} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}(sim(I_i, I_j) - sim(T_i, T_j))^2. \tag{5}$$

For the cross-modality regularization, given two image-text pairs, the gap between their distances is minimized:

$$\mathcal{L}_{C^C-ITC} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}(sim(I_i, T_j) - sim(I_j, T_i))^2. \tag{6}$$

Together, the C-ITC can be formulated as:

$$\mathcal{L}_{C-ITC} = \mathcal{L}_{C^I-ITC} + \mathcal{L}_{C^C-ITC}. \tag{7}$$

3.4 Model Generalization

Apart from the empirical study on data augmentation and loss function, we also verify the model generalization from two sides. (1) We evaluate the generalization of TBPS-CLIP as a baseline by applying it to other TBPS methods. (2) We fine-tune TBPS-CLIP using a small amount of TBPS training data to assess its generalization in the few-shot setting.

Methods	Rank-1	Rank-5	Rank-10	mAP
CLIP	60.67	81.99	88.87	54.72
CLIP + GlobalGrad	63.66	**84.08**	90.11	56.46
CLIP + Dropout	61.06	82.00	88.95	55.10
CLIP + LockBL	61.27	82.23	88.79	55.43
CLIP + SLabel	61.53	81.99	88.71	55.22
CLIP*	**64.34**	84.05	**90.51**	**57.53**

Table 1: Ablations of training tricks on CUHK-PEDES. CLIP* represents CLIP with all four training trick.

3.5 Model Compression

We provide insight into the model by evaluating the role of each module to the final performance, which is valuable for model compression in real-world applications. Two evaluation metrics (Wang and Tu 2020) are adopted. The first one evaluates a specific module's contribution by removing the module and observing the performance drop. The second one examines the module's importance by measuring how closely the module's weights can approach their initial values while maintaining a certain level of performance.

4 Experiments

The experimental analyses of the empirical studies are detailed in this section. Also, although our studies are directed at common technologies to keep the model simple, we provide insights from the TBPS-specific view in this section, *i.e.,* discussing why these technologies can work in TBPS.

Comparisons with other methods are carried out on three datasets: CUHK-PEDES (Li et al. 2017b), ICFG-PEDES (Ding et al. 2021), RSTPReid (Zhu et al. 2021). Extensive ablation studies are on CUHK-PEDES. The evaluation of performance is based on Rank-k and mean Average Precision (mAP). *The introduction of dataset, evaluation metric and implementation detail are in the Appendix.*

4.1 Ablations of Training Tricks

Starting from CLIP, we first perform experimental investigations on the training tricks discussed in Section 3.1. As shown in Table 1, all training tricks introduce a positive impact into CLIP on the performance.

4.2 Ablations of Data Augmentation

We show the optimal results of each augmentation, and analyze their effects. *The ablation study of the augmentations with different hyperparameters is shown in the Appendix.*

Image Augmentations. Table 2 studies the effect of image augmentations. (1) In the removal augmentations, RandomResizedCrop and RandomErasing enhance performance. They randomly crop/erase the part of the image, emphasizing local details and supporting cross-modal fine-grained learning in TBPS. Surprisingly, RandomGrayscale, which eliminates color information, also improves results. By inputting the augmented grayscale image into the model, the model is compelled to focus on other information like texture and shape during training. While color information is

Methods	Rank-1	Rank-5	Rank-10	mAP
CLIP*	64.34	84.05	90.51	57.53
The group of removal:				
RandomResizedCrop (✓)	65.29	85.01	90.68	58.62
RandomErasing (✓)	64.64	84.34	90.74	58.08
RandomGrayscale (✓)	65.29	84.67	90.56	58.18
GaussianBlur	55.30	77.86	85.88	49.82
The group of alternation:				
ColorJitter-BCS (✓)	64.86	84.65	90.47	57.85
ColorJitter-Hue	64.12	84.81	90.87	57.41
RandomHorizontalFlip (✓)	65.48	84.86	90.48	58.46
RandomVerticalFlip	64.23	83.98	90.16	57.35
RandomRotation (✓)	65.50	85.20	91.13	58.92
Applying multiple augmentations:				
Stacking Together	65.92	85.33	90.89	59.43
AutoAug	65.43	84.84	91.07	58.55
RandAug	65.58	**85.72**	91.00	59.08
TrivialAug	65.59	84.58	90.81	58.58
Augmentation Pool	**66.13**	85.38	**91.21**	**59.30**

Table 2: Ablations of image augmentations on CUHK-PEDES. 'Stacking Together' applies multiple removal and alternation augmentations with the '✓' symbol into the image, while 'Augmentation Pool' randomly selects two of them in each iteration. 'AutoAug', 'RandAug' and 'TrivialAug' are applied at the default setting.

known to be critical in TBPS (Wu et al. 2021; Wang et al. 2022c), these empirical results suggest that other information, apart from color, is also valuable for person retrieval. Conversely, GaussianBlur, which blurs fine-grained details, significantly degrades performance. Blurring the entire image leads to the loss of crucial fine-grained information, which is crucial for TBPS. (2) In the alternation augmentations, most are beneficial for the performance, including ColorJitter-BCS, RandomHorizontalFlip and RandomRotation. They increase image diversity without altering image semantics, enhancing the model's robustness to different images and leading to improved performance. By contrast, ColorJitter-Hue and RandomVerticalFlip hurt performance. The former alters the color, causing inconsistency between the color information in the image and its corresponding textual description, while the latter significantly alters the overall shape of the image. (3) Beyond applying the single augmentation to the image, we conduct experiments of applying multiple augmentations. Intuitively, we can stack the above effective augmentations to the image (Stacking Together), which leads to performance improvement. We also explore automatic image augmentation strategies AutoAugment (Cubuk et al. 2018), RandAugment (Cubuk et al. 2020) and TrivialAugment (Müller and Hutter 2021). Nevertheless, the proposed augmentation pool that randomly selects 2 effective augmentations each time gets the best results.

Text Augmentations. Table 3 shows that back translation and random deletion are effective text augmentations for TBPS. Back translation enriches the original texts, while random deletion acts as a regularization technique by randomly removing words. Both positively impact performance. When combined (Stacking Together), they result in a Rank-1 improvement of 1.38%. However, the synonym replacement, random insertion and random swap negatively impact performance. They risk distorting the original meaning of the text. Synonym replacement risks changing the semantics of the original text due to the errors from searching synonyms, while random insertion and random swap tend to break the sentence structure. These make the text encoder harder to comprehend the augmented text. Consequently, it is reasonable that EDA (Wei and Zou 2019), selecting one among these at random, does not enhance performance.

Methods	Rank-1	Rank-5	Rank-10	mAP
CLIP*	64.34	84.05	90.51	57.53
Back Translation (✓)	64.77	**84.67**	90.56	57.49
Synonym Replacement	63.95	83.76	89.96	56.90
Random Insertion	63.71	84.29	90.03	56.77
Random Swap	56.41	79.08	86.14	49.29
Random Deletion (✓)	65.53	84.70	**90.87**	57.90
EDA	63.43	83.77	90.27	56.15
Stacking Together	**65.72**	84.62	90.76	**58.68**

Table 3: Ablations of text augmentations on CUHK-PEDES. 'Stacking Together' applies multiple augmentations with the '✓' symbol into the text.

Methods	Rank-1	Rank-5	Rank-10	mAP
CLIP*	64.34	84.05	90.51	57.53
Image Aug	66.13	85.38	91.21	59.30
Text Aug	65.72	84.62	90.76	58.68
Image & Text Aug	**66.78**	**85.98**	**91.23**	**59.68**

Table 4: Results of combining optimal augmentations within each modality.

Together with Optimal Augmentations. As shown in Table 4, based on the CLIP*, after combining all optimal data augmentations (*i.e.*, 'Augmentation Pool' for image, 'Stacking Together' for text), the Rank-1 accuracy is significantly increased by 2.44%. It is worth stressing that the gain is only from exploiting data augmentations.

4.3 Ablations of Loss Functions

Table 5 studies the effectiveness of the loss functions. (1) Replacing the original image-text contrastive loss with the normalized image-text contrastive loss (N-ITC) in CLIP*+Aug results in a slight improvement of 0.13% at Rank-1. (2) Among the loss functions aimed at improving data efficiency, image self-supervision (SS-I) achieves the best performance compared to SS-T and SS-IT. Similarly, multi-view supervision of image (MVS-I) outperforms MVS-T and MVS-IT. These findings indicate that leveraging image data for improving data efficiency is more effec-

Methods	Rank-1	Rank-5	Rank-10	mAP
CLIP* +Aug	66.78	85.98	91.23	59.68
N-ITC (✓)	66.91	85.98	91.68	60.01
Improving data efficiency:				
SS-I (✓)	67.54	86.13	**91.78**	60.94
SS-T	65.97	85.40	90.72	59.10
SS-IT	66.81	86.09	91.63	59.82
MVS-I (✓)	67.50	86.45	91.63	60.84
MVS-T	66.46	85.96	91.13	60.24
MVS-IT	67.01	85.66	91.33	59.83
Optimization for retrieval:				
R-ITC (✓)	68.19	86.01	90.94	60.90
C-ITC (✓)	67.15	86.31	92.04	60.58
Applying multiple losses:				
Stacking Together (TBPS-CLIP)	**69.54**	**86.99**	91.24	**61.57**

Table 5: Ablations of loss functions on CUHK-PEDES. 'Stacking Together' denotes that multiple losses with the '✓' symbol are used to CLIP*+Aug .

Methods	Rank-1	Rank-5	Rank-10	mAP
w/o CLIP:				
ViTAA (Wang et al. 2020)	55.97	75.84	83.52	-
SSAN (Ding et al. 2021)	61.37	80.15	86.73	-
LapsCore (Wu et al. 2021)	63.40	-	87.80	-
LGUR (Shao et al. 2022)	65.25	83.12	89.00	-
SAF (Li, Cao, and Zhang 2022)	64.13	82.62	88.40	58.61
IVT (Shu et al. 2023)	65.59	83.11	89.21	60.66
RaSa (Bai et al. 2023a)	**76.51**	**90.29**	**94.25**	**69.38**
w/ CLIP:				
TBPS-LD (Han et al. 2021)	64.08	81.73	88.19	60.08
CFine (Yan et al. 2022)	69.57	85.93	91.15	-
TP-TPS (Wang et al. 2023)	70.16	86.10	90.98	66.32
IRRA (Jiang and Ye 2023)	73.38	**89.93**	**93.71**	**66.13**
CLIP (ViT-B/32)	60.67	81.99	88.87	54.72
TBPS-CLIP (ViT-B/32)	69.54	86.99	91.24	61.57
CLIP (ViT-B/16)	65.37	85.83	91.59	59.42
TBPS-CLIP (ViT-B/16)	**73.54**	88.19	92.35	65.38
Simplified TBPS-CLIP (ViT-B/16)	72.66	88.14	92.72	64.97

Table 6: Comparison with other methods on CUHK-PEDES.

tive than using text data in TBPS. (3) Both R-ITC and C-ITC, which are loss functions optimized for retrieval, improve performance. Notably, R-ITC leads to a significant boost of 1.41% compared to CLIP+Aug, highlighting the importance of R-ITC's constraint in pulling negative samples apart. Finally, we combine all of these effective losses, boosting performance by a large margin of 2.76% at Rank-1 based on CLIP*+Aug, achieving 69.54% Rank-1 accuracy.

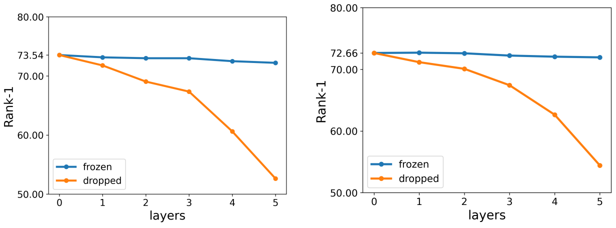
4.4 Comparisons with State-of-the-Art Methods

We compare TBPS-CLIP with other methods on three datasets in Tables 6-8. *We provide a visualization of the comparison of retrieval results in the Appendix.*

(1) Compared to methods using CLIP, TBPS-CLIP with ViT-B/16 as the image encoder outperforms the state-of-the-art method IRRA on ICFG-PEDES and RSTPReid, while achieving similar performance on CUHK-PEDES. Notably, TBPS-CLIP maintains a simple network architecture, unlike IRRA which incorporates a multimodal interaction encoder after CLIP. Despite its simplicity, TBPS-CLIP achieves promising results and is more lightweight than IRRA, as shown in Table 9. Its efficient training allows for training in just 5 epochs, making it a friendly baseline. (2) Compared with the methods without CLIP, we notice that RaSa performs strongly. It adopts ALBEF (Li et al. 2021a) as the baseline and contains two models, an online model and its momentum version, each consisting of an image encoder, a text encoder and a cross-modal encoder. Although having outstanding performance, RaSa is cumbersome and challenging to support its generalization, as shown in Table 9. The proposed TBPS-CLIP, with very lightweight and low-cost architecture and promising performance, has the potential as a baseline to provide broader applications. (3) Considering the convenience of applying TBPS-CLIP as the baseline, we further provide a simplified TBPS-CLIP, in which there are only N-ITC and R-ITC losses. It still performs satisfactorily, specifically, even beats IRRA on ICFG-PEDES and RSTPReid. The simplified TBPS-CLIP equipped with two losses can be more easily applied as a baseline in further work.

4.5 More Probing Experiments of TBPS-CLIP

Model Generalization. (1) We select IRRA (Jiang and Ye 2023), the most advanced CLIP-based TBPS method so far, as the hotbed for verifying the generalization and effectiveness of TBPS-CLIP baseline. Specifically, we adopt TBPS-CLIP (ViT-B/16) and its simplified version as the IRRA's baseline, respectively, instead of the original CLIP (ViT-B/16). Table 10 demonstrates the generalization and effectiveness of TBPS-CLIP. *More experimental results on other datasets are shown in the Appendix.* (2) We study the few-shot capabilities (5% training data) of TBPS-CLIP in Table 11, and *more experimental results with 1% training data and 10% one are shown in the Appendix.* CLIP presents a poor performance in few-shot TBPS, especially, the representative few-shot CLIP variants (CoOp and CLIP-



(a) TBPS-CLIP (b) Simplified TBPS-CLIP

Figure 3: The trend of performance when dropping/freezing some layers of the text encoder on CUHK-PEDES.

Methods	Rank-1	Rank-5	Rank-10	mAP
w/o CLIP:				
ViTAA (Wang et al. 2020)	50.98	68.79	75.78	-
SSAN (Ding et al. 2021)	54.23	72.63	79.53	-
LGUR (Shao et al. 2022)	57.42	74.97	81.45	-
SAF (Li, Cao, and Zhang 2022)	54.86	72.13	79.13	32.76
IVT (Shu et al. 2023)	56.04	73.60	80.22	-
RaSa (Bai et al. 2023a)	**65.28**	**80.40**	**85.12**	**41.29**
w/ CLIP:				
TP-TPS (Wang et al. 2023)	60.64	75.97	81.76	**42.78**
CFine (Yan et al. 2022)	60.83	76.55	82.42	-
IRRA (Jiang and Ye 2023)	63.46	80.25	**85.82**	38.06
CLIP (ViT-B/32)	53.96	73.69	80.43	32.37
TBPS-CLIP (ViT-B/32)	59.88	77.40	83.33	34.96
CLIP (ViT-B/16)	55.97	74.62	81.35	30.63
TBPS-CLIP (ViT-B/16)	**65.05**	**80.34**	85.47	39.83
Simplified TBPS-CLIP (ViT-B/16)	64.52	80.03	85.39	39.54

Table 7: Comparison with other methods on ICFG-PEDES.

Methods	Rank-1	Rank-5	Rank-10	mAP
w/o CLIP:				
SSAN (Ding et al. 2021)	43.50	67.80	77.15	-
SAF (Li, Cao, and Zhang 2022)	44.05	67.30	76.25	36.81
IVT (Shu et al. 2023)	46.70	70.00	78.80	-
RaSa (Bai et al. 2023a)	**66.90**	**86.50**	**91.35**	**52.31**
w/ CLIP:				
CFine (Yan et al. 2022)	50.55	72.50	81.60	-
TP-TPS (Wang et al. 2023)	50.65	72.45	81.20	43.11
IRRA (Jiang and Ye 2023)	60.20	81.30	88.20	47.17
CLIP (ViT-B/32)	50.10	76.10	84.95	41.14
TBPS-CLIP (ViT-B/32)	56.65	80.75	87.30	44.00
CLIP (ViT-B/16)	56.15	78.30	86.60	43.26
TBPS-CLIP (ViT-B/16)	61.95	**83.55**	**88.75**	**48.26**
Simplified TBPS-CLIP (ViT-B/16)	**62.10**	81.90	87.75	48.00

Table 8: Comparison with other methods on RSTPReid.

Methods	Baselines	Param.(M)	Epoch	Time(s)	
				Training	Test
RaSa	ALBEF	210.2	30	27967.5	869.8
IRRA	CLIP (ViT-B/16)	194.5	60	6110.4	31.4
TBPS-CLIP	CLIP (ViT-B/16)	149.2	5	1234.7	31.4

Table 9: Comparisons on CUHK-PEDES. Param.(M) and Epoch denote the number of modal parameters (in millions) and the number of epochs in training, respectively. Time(s) represents the online running time (in seconds).

Methods	Baselines	Rank-1	Rank-5	Rank-10	mAP	mINP
IRRA	CLIP	73.38	89.93	93.71	66.13	50.24
IRRA*	TBPS-CLIP	74.97	89.82	93.80	67.84	52.53
IRRA*	S. TBPS-CLIP	74.56	89.26	93.52	67.52	52.58

Table 10: Results of using TBPS-CLIP as baseline on CUHK-PEDES. The mean Inverse Negative Penalty (mINP) is used in IRRA and also adopted here for comparison. 'S. TBPS-CLIP' is the abbreviation for simplified TBPS-CLIP.

Methods	Rank-1	Rank-5	Rank-10	mAP
CLIP (Radford et al. 2021)	38.37	62.26	72.34	35.39
CoOp (Zhou et al. 2022)	11.37	24.19	32.46	10.48
CLIP-Adapter (Gao et al. 2021)	11.96	25.45	33.56	10.91
TBPS-CLIP	42.98	66.26	74.94	38.86
S. TBPS-CLIP	43.65	66.60	75.91	39.30

Table 11: Performance under few-shot settings (5% training data) on CUHK-PEDES.

Adapter) are even worse in performance. The three methods are skilled in the few-shot image classification since the large-scale data knowledge of CLIP from the pre-training phase shares the same data characteristics as the downstream classification task. However, TBPS is a fine-grained person-specific task and has a noticeable gap with the pre-trained data in CLIP. The TBPS performance will be poor if there is insufficient training data to fine-tune CLIP, and CoOp and CLIP-Adapter with the locked CLIP backbone in the fine-tuning phase also perform poorly in TBPS. Alternatively, the proposed TBPS-CLIP with powerful learning capacity in TBPS can alleviate these problems and brings promising few-shot results. More than that, the simplified TBPS-CLIP has superiority in the few-shot setting.

Model Compression. We compute the two metrics of TBPS-CLIP in Section 3.5 as the guidance for the model compression. *The computation of the metrics, along with the investigation of TBPS-CLIP's internal properties and a detailed explanation of model compression, are provided in the Appendix.* Finally, from Figure 3, we can clearly see that freezing some layers of the text encoder does not have much impact on performance while the dropping operation can negatively affect performance. As a result, we can compress TBPS-CLIP during training by freezing part of text encoder.

5 Conclusion

This paper makes a thorough empirical study to explore the potential of CLIP for TBPS. We empirically prove that CLIP, only equipped with common data augmentations, loss functions, and practical training tricks (without complex modules), can achieve promising results on multiple TBPS benchmarks. Further, we prove the effectiveness of the proposed TBPS-CLIP in terms of model generalization and model compression. Our empirical study aims to offer practical guidance for future research on CLIP-based TBPS.

Acknowledgments

This work was supported by the National Science Foundation of China under Grant NSFC 62002252, NSFC 62066021 and the Priority Academic Program Development of Jiangsu Higher Education Institutions.

References

Bai, Y.; Cao, M.; Gao, D.; Cao, Z.; Chen, C.; Fan, Z.; Nie, L.; and Zhang, M. 2023a. RaSa: Relation and Sensitivity Aware Representation Learning for Text-based Person Search. *arXiv preprint arXiv:2305.13653*.

Bai, Y.; Wang, J.; Cao, M.; Chen, C.; Cao, Z.; Nie, L.; and Zhang, M. 2023b. Text-based Person Search without Parallel Image-Text Data. *arXiv preprint arXiv:2305.12964*.

Cao, M.; Li, S.; Li, J.; Nie, L.; and Zhang, M. 2022. Image-text retrieval: A survey on recent research and development. *arXiv preprint arXiv:2203.14713*.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PMLR.

Chen, X.; and He, K. 2021. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 15750–15758.

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

Cubuk, E. D.; Zoph, B.; Mane, D.; Vasudevan, V.; and Le, Q. V. 2018. Autoaugment: Learning augmentation policies from data. *arXiv preprint arXiv:1805.09501*.

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

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*.

Gao, P.; Geng, S.; Zhang, R.; Ma, T.; Fang, R.; Zhang, Y.; Li, H.; and Qiao, Y. 2021. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544*.

Goel, S.; Bansal, H.; Bhatia, S.; Rossi, R.; Vinay, V.; and Grover, A. 2022. Cyclip: Cyclic contrastive language-image pretraining. *Advances in Neural Information Processing Systems*, 35: 6704–6719.

Han, X.; He, S.; Zhang, L.; and Xiang, T. 2021. Text-based person search with limited data. *arXiv preprint arXiv:2110.10807*.

Jiang, D.; and Ye, M. 2023. Cross-Modal Implicit Relation Reasoning and Aligning for Text-to-Image Person Retrieval. *arXiv preprint arXiv:2303.12501*.

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

Li, J.; Selvaraju, R.; Gotmare, A.; Joty, S.; Xiong, C.; and Hoi, S. C. H. 2021a. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34: 9694–9705.

Li, S.; Cao, M.; and Zhang, M. 2022. Learning semantic-aligned feature representation for text-based person search. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2724–2728. IEEE.

Li, S.; Xiao, T.; Li, H.; Yang, W.; and Wang, X. 2017a. Identity-aware textual-visual matching with latent co-attention. In *Proceedings of the IEEE International Conference on Computer Vision*, 1890–1899.

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

Li, Y.; Liang, F.; Zhao, L.; Cui, Y.; Ouyang, W.; Shao, J.; Yu, F.; and Yan, J. 2021b. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *arXiv preprint arXiv:2110.05208*.

Lu, J.; Batra, D.; Parikh, D.; and Lee, S. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

Mu, N.; Kirillov, A.; Wagner, D.; and Xie, S. 2022. Slip: Self-supervision meets language-image pre-training. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVI*, 529–544. Springer.

Müller, S. G.; and Hutter, F. 2021. Trivialaugment: Tuning-free yet state-of-the-art data augmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 774–782.

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

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

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

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

Wang, C.; Chai, M.; He, M.; Chen, D.; and Liao, J. 2022a. Clip-nerf: Text-and-image driven manipulation of neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3835–3844.

Wang, G.; Yu, F.; Li, J.; Jia, Q.; and Ding, S. 2023. Exploiting the Textual Potential from Vision-Language Pre-training for Text-based Person Search. *arXiv preprint arXiv:2303.04497*.

Wang, W.; and Tu, Z. 2020. Rethinking the value of transformer components. *arXiv preprint arXiv:2011.03803*.

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

Wang, Z.; Lu, Y.; Li, Q.; Tao, X.; Guo, Y.; Gong, M.; and Liu, T. 2022b. Cris: Clip-driven referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11686–11695.

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

Wei, J.; and Zou, K. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*.

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

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

Yao, L.; Huang, R.; Hou, L.; Lu, G.; Niu, M.; Xu, H.; Liang, X.; Li, Z.; Jiang, X.; and Xu, C. 2021. FILIP: fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*.

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.

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

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

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9): 2337–2348.

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.