# Clean Sample Selection and Noisy Sample Rematching for Text-Based Pedestrian Retrieval

Daiqiang Li[1]
lidaiqiang@stu.scu.edu.cn
Weicheng Zhang[1]
weicheng@stu.scu.edu.cn
Yuanyuan Wu[2]
wuyuanyuan@cdut.edu.cn
Honggang Chen[†1,3]
honggang_chen@scu.edu.cn

[1] School of Electronics and Information Engineering, Sichuan University, Chengdu, China
[2] School of Computer Science and Cyber Security, Chengdu University of Technology, Chengdu, China
[3] Yunnan Key Laboratory of Software Engineering, Yunnan University, Kunming, China

**Abstract**

Text-Based Pedestrian Retrieval (TBPR) aims to match pedestrian images with corresponding text descriptions. Although notable advancements have been made in this field, most existing methods assume that training samples are clean. However, the neglect of noisy samples can degrade model performance. To address this issue, we propose an effective framework called CSNR from the perspective of dynamic dataset reconstruction, incorporating three key innovations. The Clean Sample Selection Module (CSSM) exploits the differences in the loss values of clean and noisy samples, filtering out noisy samples to improve the quality of training samples. The Noisy Sample Rematching Module (NSRM) replaces the text descriptions of some noisy samples with better-matching descriptions to create new clean samples, thereby increasing the set of usable clean data. The Weighted topK Contrastive Loss (WKCL) uses more negative samples and prioritizes harder negatives to fully leverage the information from negative samples and ensure the efficiency of model training under noisy description circumstances. Experimental results demonstrate that our approach achieves state-of-the-art performance.

## 1 Introduction

Text-Based Pedestrian Retrieval (TBPR) aims to match a given text description with the corresponding image [27, 30]. In recent years, deep learning-based TBPR methods have shown remarkable progress. Some of them [12, 14] are dedicated to addressing cross-modal alignment to overcome significant intra-class variations. Several methods [5, 15, 29] focus on the identity-level matching. Other works [10, 11, 26] are committed to resolving the occlusion problem. A recent study [20] indicates that most methods assume that all samples used for model training are clean. However, the noisy description problem, *i.e.*, the misalignment

[†] Corresponding author

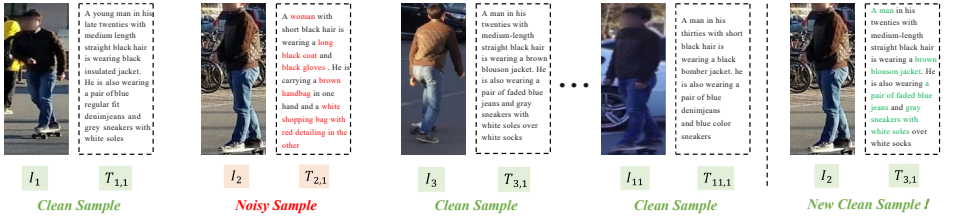


Figure 1: The corresponding image and text descriptions of the same person from the ICFG-PEDES [4] dataset. The second image $I_2$ does not align with its corresponding description $T_{2,1}$ and can thus be considered a noisy sample. Notably, the third description $T_{3,1}$ perfectly matches the second image. Therefore, we can create a new clean sample $(I_2, T_{3,1})$ by replacing the noisy sample's description.

between text descriptions and images, as shown in Figure 1, exists in the training dataset. This misalignment in noisy samples can seriously undermine model performance [8], but it is often overlooked in TBPR.

To address the noisy description problem in TBPR, we propose an effective framework, consisting of the Clean Sample Selection Module (CSSM), the Noisy Sample Rematching Module (NSRM), and the Weighted topK Contrastive Loss (WKCL), collectively referred to as CSNR. According to the memorization effect of DNNs [1], the average loss value for clean samples tends to be lower than that for noisy samples. Based on this effect, CSSM is proposed to divide the dataset into clean and noisy subsets through a loss comparison method. Besides, as shown in Figure 1, datasets for TBPR generally contain multiple images per individual, and each image is associated with one or two descriptions. It is worth mentioning that a single description can describe multiple images of the same individual, beyond its original match. Taking advantage of this characteristic of the dataset, the NSRM converts some noisy samples to clean ones by rematching other text descriptions to the images of the noisy samples from the same person, thus expanding the set of usable clean samples. Additionally, we propose the WKCL, which considers more negative samples and prioritizes harder negatives, compared to the traditional Triplet loss [22] and E-loss [25]. WKCL effectively leverages the information from negative samples and improves the efficiency of model training under noisy description circumstances.

To conclude, the key contributions of our work can be summarized as follows:

- To address the noisy description problem in TBPR, we propose an effective framework from the perspective of dynamic dataset reconstruction in the training process, including CSSM, NSRM, and WKCL. Extensive experiments demonstrate the effectiveness of our framework.

- The CSSM filters out noisy samples by leveraging the differences in loss values between clean and noisy samples, thereby improving the quality of the training samples.

- The NSRM, designed based on the characteristics of the dataset, aims to transform some noisy samples into clean samples by replacing their descriptions with more appropriate ones from the same person. This effectively expands the set of usable training samples.

- The WKCL leverages more negative samples while emphasizing harder negatives during training, effectively utilizing the information from negative samples and improving the efficiency of model training under noisy description circumstances.

# 2 Related Work

## 2.1 Text-Based Pedestrian Retrieval

Text-based pedestrian retrieval aims to match the image of a person to the natural-language description. The technique has become increasingly important in criminal case investigation and security applications. Early works typically rely on separately pretrained unimodal encoders, such as ResNet-50/101 [6] for vision and BERT [3] for language, to extract visual and textual features. However, handling the two modalities in isolation inevitably leaves an alignment gap that hinders precise cross-modal matching. Recently, VLP (Vision-Language Pre-training) largely improves model performance in cross-modal alignment. In particular, CLIP [21] has become the mainstream backbone for TBPR by jointly learning visual and textual representations and thus overcoming the limitations of single-modality pre-training. Nonetheless, most existing TBPR methods ignore the pervasive noisy samples in datasets, which markedly degrade retrieval performance.

## 2.2 Methods for Noisy Description Problem

In the image-text matching field, there is also the problem of noisy description, and several methods have been proposed to tackle this issue. Based on the memorization effect of DNNs, many approaches [8, 13] first compute the per-sample loss, then model these losses using a Gaussian Mixture Model (GMM) [19] to estimate the probability of each sample belonging to the clean set, followed by subsequent processing steps. Recently, the Beta Mixture Model [31] is proposed, demonstrating superior performance in categorization. To mitigate the selection bias of clean samples, Co-teaching methods [33] are proposed to train the different networks with small loss values. These existing methods for the noisy description problem in the image-text matching field can provide some reference. However, they fail to leverage the characteristics of datasets in the TBPR reasonably. Additionally, the commonly used contrastive loss functions, such as the Triplet loss and the E-loss, are not suitable for the processing strategy of negative samples under noisy description circumstances, resulting in performance degradation. Thus, it is significant to address the noisy description problem in TBPR.

# 3 Method

We assume that $\left(I_i^n, T_{i,j}^n\right)$ represents one sample pair of person $n$, where $n \in \{1,2,3,\ldots,N\}$, and $N$ is the total number of individuals in the training dataset. $i \in \{1,2,3,\ldots,N_n\}$, where $N_n$ denotes the total number of image samples for person $n$. $j \in \{1,2\}$, as each image may have one or two corresponding descriptions, depending on the specific dataset. For convenience, we omit the superscript $n$ when distinguishing individuals is unnecessary in some subsequent cases. Figure 1 shows some samples presented in this manner.

The presence of noisy samples degrades model performance. To address this problem, we propose a framework called CSNR, as illustrated in Figure 2. Firstly, we extract the vision

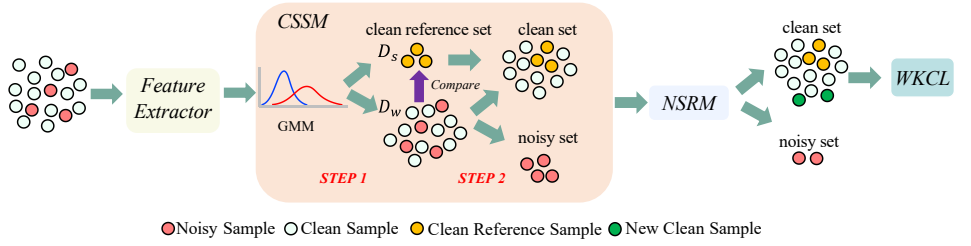


Figure 2: The overview of our framework. We reconstruct the dataset dynamically through CSSM and NSRM, improving the quality of training samples and increasing the set of usable clean samples. The model is trained on the clean samples using WKCL.

and textual features using the Cross-Modal Feature Extractor. Then, WKCL is employed to calculate the per-sample loss, which will be fed into the GMM. In the first step of CSSM, we select some samples as clean reference samples by GMM. After that, the remaining samples are split into the clean set and the noisy set based on the loss comparison with the corresponding clean reference samples. After that, the NSRM converts some noisy samples into new clean samples. Finally, the model is trained on clean samples with WKCL.

## 3.1 Cross-Modal Feature Extractor

We adopt CLIP [21] as the backbone for feature extraction, as shown in Figure 3(a). By inputting $(I_i, T_{i,j})$ into the image encoder $g^v$ and the text encoder $g^t$, we obtain the token sequences $\{v_i^{cls}, v_i^1, v_i^2, \ldots, v_i^{N_v}\}$ and $\{t_{i,j}^{sos}, t_{i,j}^1, t_{i,j}^2, \ldots, t_{i,j}^{N_t}, t_{i,j}^{eos}\}$, where $N_v+1$ and $N_t+2$ denote the lengths of the image and text token sequence, respectively. $\hat{V}_i = v_i^{cls}$ and $\hat{T}_{i,j} = t_{i,j}^{eos}$ represent the coarse-grained token, while the other tokens represent the fine-grained tokens. The coarse-grained token is considered the coarse-grained features (CF). As for the fine-grained features (FF), following the previous work [20], we first select some effective tokens that are more informative to represent the fine-grained tokens. It is obvious that the self-attention map $A_v \in \mathbb{R}^{(N_v+1)(N_v+1)}$ for the image and $A_t \in \mathbb{R}^{(N_t+2)(N_t+2)}$ for the text description comprise the correlation between the fine-grained tokens and the coarse-grained tokens, *i.e.*, $A_v[0, 1:N_v+1]$ and $A_t[0, 1:N_t+1]$. We select $M\%$ tokens with higher correlation to the coarse-grained features as effective tokens which are subsequently input into a Multilayer Perceptron, followed by a fully connected layer and a max pooling layer to obtain the fine-grained features $\check{V}_i$ and $\check{T}_{i,j}$. Coarse-grained features and fine-grained features represent global information and local information, respectively. They are used to calculate the similarity and the loss value comprehensively.

## 3.2 Clean Sample Selection Module (CSSM)

We propose the CSSM, as illustrated in Figure 3(b), to filter out noisy samples from the dataset during each training epoch. According to the DNNs' memorization effect [1], the mean loss value of clean samples is smaller than that of noisy samples. Therefore, we compute the per-sample loss of every sample by the coarse-grained and fine-grained features using WKCL, which will be introduced in detail later. To obtain the probability of each sample being clean, we fed all the per-sample losses into the two-component Gaussian Mixture

Model (GMM), namely

$$P(u|\theta) = \sum_{k=1}^{K} \alpha_k \phi(u|\theta_k) \tag{1}$$

where $\alpha_k$ represents the mixture coefficient, $k = 1$ denotes the clean set and $k = 2$ denotes the noisy set. The GMM is optimized with the Expectation-Maximization algorithm. After that, the probability of the sample pair $(I_i, T_{i,j})$ being clean can be computed as follow:

$$p(\theta_1|u_{i,j}) = p(\theta_1)p(u_{i,j}|\theta_1)/p(u_{i,j}) \tag{2}$$

where $u_{i,j}$ denotes $L_{wkcl}(I_i, T_{i,j})$, *i.e.*, the per-sample loss of $(I_i, T_{i,j})$. If we directly use $p(\theta_1|u_{i,j})$ to classify the dataset into clean and noisy sample subsets, many noisy samples can be misclassified as clean samples [34], resulting in poor classification performance.

Therefore, we propose an effective two-stage clean sample selection method by comparing the loss values. Our method is founded on a reasonable assumption: For two samples $(I_a, T_{a,j})$ and $(I_b, T_{b,j})$ with similar but different images, if they are both clean samples, their loss values should not differ significantly from each other. If $(I_a, T_{a,j})$ is clean and $(I_b, T_{b,j})$ is noisy, $L_{wkcl}(I_a, T_{a,j})$ should be smaller than $L_{wkcl}(I_b, T_{b,j})$. Guided by this assumption, the dataset is classified into two subsets in the first stage: a clean reference subset $D_s = \{(I_i, T_{i,j}) \mid p(\theta_1|u_{i,j}) > 0.9\}$, which consists of a small number of samples that are highly likely to be clean, and a subset $D_w$, which includes the remaining samples that require further evaluation in the second stage.

Next, for a sample $(I_k, T_{k,j}) \in D_w$, we find the reference sample $(I_q, T_{q,j}) \in D_s$ whose image $I_q$ is the most similar image to $I_k$ among the images in $D_s$. The similarity is measured from both coarse-grained and fine-grained perspectives. Specifically, the similarity $R(I_k, I_q)$ between $I_k$ and $I_q$ equals to $(S(\hat{V}_k, \hat{V}_q) + S(\check{V}_k, \check{V}_q))/2$, where $S(m,n) = m^\top n / \|m\| \|n\|$ represent the cosine similarity between $m$ and $n$. To ensure a comprehensive evaluation, we compare the loss values calculated using coarse-grained and fine-grained features separately. The loss function will be introduced later. If the following conditions can be satisfied, $(I_k, T_{k,j})$ is considered a clean sample.

$$\begin{cases} \hat{L}_{wkcl}(I_k, T_{k,j}) - \hat{L}_{wkcl}(I_q, T_{q,j}) \leq \mu \\ \check{L}_{wkcl}(I_k, T_{k,j}) - \check{L}_{wkcl}(I_q, T_{q,j}) \leq \mu \end{cases} \tag{3}$$

where $\mu$ is a hyperparameter which we will analyze later in the experiment section. The samples in the $D_w$ which satisfied the mentioned conditions and samples in the $D_s$ are considered clean samples altogether. The remaining samples in the $D_w$ are considered noisy samples.

## 3.3 Noisy Sample Rematching Module (NSRM)

As presented in Figure 3(c), NSRM converts some noisy samples into clean samples by rematching other appropriate descriptions from the same person. For the $n$-th person, if sample $(I_i^n, T_{i,j}^n)$ is a noisy sample, we calculate the cosine similarity between $I_i^n$ and every description $T_{k,j}^n$ $(k \neq i)$ from the same person which belongs to a clean sample. The similarity $R(I_i^n, T_{k,j}^n)$ between $I_i^n$ and $T_{k,j}^n$ equals to $(S(\hat{V}_i^n, \hat{T}_{k,j}^n) + S(\check{V}_i^n, \check{T}_{k,j}^n))/2$. We choose the most similar $T_{k,j}^n$ to be the candidate description. If $R(I_i^n, T_{k,j}^n)$ satisfies the following conditions, $T_{k,j}^n$ will replace the original description $T_{i,j}^n$:

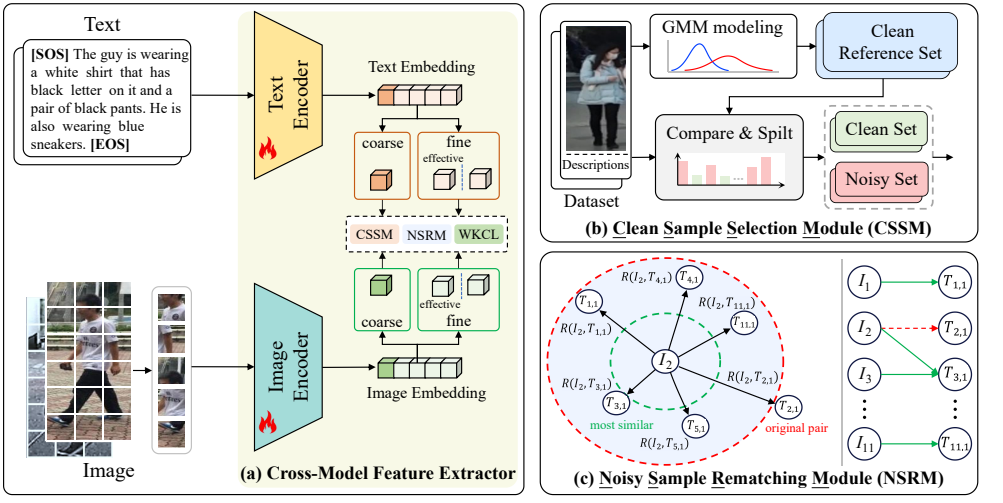


Figure 3: (a) The Cross-Modal Feature Extractor can extract both coarse-grained and fine-grained features from images and descriptions. (b) The CSSM builds a clean reference set in the first stage, and then in the second stage, it distinguishes clean samples from noisy samples by comparing loss values. (c) The NSRM makes full use of the characteristics of the dataset in TBPR by replacing the descriptions of noisy samples with other more appropriate descriptions of the same individual, thereby reconstructing clean samples.

- The similarity $R(I_i^n, T_{k,j}^n)$ between $I_i^n$ and candidate description $T_{k,j}^n$ is larger than the similarity $R(I_i^n, T_{i,j}^n)$ between $I_i^n$ and $T_{i,j}^n$.
- We calculate the similarity between the image of every noisy sample in the dataset and their candidate descriptions. $R(I_i^n, T_{k,j}^n)$ can be the top $\eta\%$.

$\eta\%$ is a hyperparameter which will be further analyzed later. These two principles ensure the replacement can effectively convert some noisy samples into new clean samples. At least, the candidate description is more appropriate than the original description. Finally, the model is trained on the clean samples classified by the CSSM and the new clean samples rematched by the NSRM.

## 3.4 Weighted topK Contrastive Loss (WKCL)

Triplet loss relies solely on the hardest negatives, disregarding other negative samples, which makes it less effective under noisy description circumstances. Meanwhile, E-loss lacks sufficient emphasis on harder negatives, resulting in suboptimal training efficiency. To overcome these limitations, we propose WKCL. For the sample $(I_i, T_{i,j})$, the WKCL is composed of the coarse-grained loss (CL) $\hat{L}_{wkcl}(I_i, T_{i,j})$ and the fine-grained loss (FL) $\check{L}_{wkcl}(I_i, T_{i,j})$, which are calculated based on the coarse-grained features and fine-grained features, respectively. To comprehensively evaluate the similarity, both CL and FL include two parts, which are from the perspective of image to description and description to image. Altogether, the WKCL is defined as:

$$\begin{aligned} L_{wkcl}(I_i, T_{i,j}) =& \hat{L}_{wkcl}(I_i, T_{i,j})_{i2t} + \hat{L}_{wkcl}(I_i, T_{i,j})_{t2i} \\ &+ \check{L}_{wkcl}(I_i, T_{i,j})_{i2t} + \check{L}_{wkcl}(I_i, T_{i,j})_{t2i} \end{aligned} \tag{4}$$

Since $CL = \hat{L}_{wkcl}(I_i, T_{i,j})_{i2t} + \hat{L}_{wkcl}(I_i, T_{i,j})_{t2i}$ and $FL = \check{L}_{wkcl}(I_i, T_{i,j})_{i2t} + \check{L}_{wkcl}(I_i, T_{i,j})_{t2i}$ are similar in Eq. (4), we only provide the definition of the CL as follows:

$$\hat{L}_{wkcl}(I_i, T_{i,j})_{i2t} = \max\left(\alpha - \hat{P}(I_i) + \hat{N}(I_i), 0\right) \tag{5}$$

$$\hat{L}_{wkcl}(I_i, T_{i,j})_{t2i} = \max\left(\alpha - \hat{P}(T_{i,j}) + \hat{N}(T_{i,j}), 0\right), \tag{6}$$

where $\alpha$ denotes the positive margin coefficient, $\hat{P}(I_i) = \hat{P}(T_{i,j}) = S(\hat{V}_i, \hat{T}_{i,j})$ denotes the cosine similarity with the corresponding positive description and image. $\hat{N}(I_i)$ and $\hat{N}(T_{i,j})$ denote the weighted similarity between the topK negative descriptions and images.

Specifically, for $N(I_i)$, we first calculate the cosine similarity between $I_i$ and every other text description of negative sample in the minibatch. Then we select the coarse-grained features of the descriptions with top $K\%$ largest similarities, *i.e.*, the descriptions of top $K\%$ hardest negatives to get the negative sequences $\mathbf{G}(I_i)$. $N(T_{i,j})$ is calculated in the same way and get $\mathbf{G}(T_{i,j})$. Specifically,

$$\hat{N}(I_i) = \sum_{k \neq i,j} \frac{e^{\frac{S(\hat{V}_i, \hat{T}_{k,j})}{\tau}}}{\sum\limits_{k \neq i,j} e^{\frac{S(\hat{V}_i, \hat{T}_{k,j})}{\tau}}} S(\hat{V}_i, \hat{T}_{k,j}), \quad \hat{N}(T_{i,j}) = \sum_{k \neq i} \frac{e^{\frac{S(\hat{V}_k, \hat{T}_{i,j})}{\tau}}}{\sum\limits_{k \neq i} e^{\frac{S(\hat{V}_k, \hat{T}_{i,j})}{\tau}}} S(\hat{V}_k, \hat{T}_{i,j}) \tag{7}$$

where $\hat{T}_{k,j} \in \mathbf{G}(I_i)$ and $\hat{V}_k \in \mathbf{G}(T_{i,j})$, $\tau$ is the temperature coefficient that controls the hardness. The softmax operation in Eq. (7) assigns greater weights to the harder negatives. To better explain why WKCL can make model training more effective under the noisy description circumstances, we conduct the gradient analysis in the Appendix.

Table 1: Performance comparisons with the state-of-the-art methods.

| Method | CUHK-PEDES | | | | ICFG-PEDES | | | | RSTPReid | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Rank-1 | Rank-5 | Rank-10 | mAP | Rank-1 | Rank-5 | Rank-10 | mAP | Rank-1 | Rank-5 | Rank-10 | mAP |
| CFine (TIP'23) [23] | 69.57 | 85.93 | 91.15 | - | 60.83 | 76.55 | 82.42 | - | 50.55 | 72.50 | 81.60 | - |
| IRRA (CVPR'23) [9] | 73.38 | 89.93 | 93.71 | 66.13 | 63.46 | 80.25 | 85.82 | 38.06 | 60.20 | 81.30 | 88.20 | 47.17 |
| VGSG (TIP'23) [8] | 71.38 | 86.75 | 91.86 | **67.91** | 63.05 | 78.43 | 84.36 | - | - | - | - | - |
| CFFKT (SPL'24) [30] | 73.91 | - | 93.91 | 66.56 | 64.22 | - | 85.54 | 39.28 | 60.95 | - | 88.55 | 47.67 |
| FSRL (ICMR'24) [27] | 74.65 | 89.77 | 94.03 | 67.49 | 64.01 | 80.42 | 85.86 | 39.64 | 60.20 | 81.40 | 88.60 | 47.38 |
| LSPM (TMM'24) [20] | 74.38 | 89.51 | 93.42 | 67.74 | 64.40 | 79.96 | 85.41 | **42.60** | - | - | - | - |
| IRLT (AAAI'24) [13] | 74.46 | **90.19** | 94.01 | - | 64.72 | 81.35 | 86.31 | - | 61.49 | 82.26 | 89.23 | - |
| CFAM (CVPR'24) [36] | 73.67 | 89.71 | 93.57 | 65.94 | 63.57 | 80.57 | 86.32 | 38.34 | 60.51 | 82.85 | 89.71 | 47.64 |
| TBPS-CLIP (AAAI'24) [6] | 73.54 | 88.19 | 92.35 | 65.38 | 65.05 | 80.34 | 85.47 | 39.83 | 61.95 | 83.55 | 88.75 | 48.26 |
| MACF (IJCV'24) [25] | 73.33 | 88.57 | 93.02 | - | 62.95 | 79.93 | 85.04 | - | - | - | - | - |
| DM-Adapter (AAAI'25) [10] | 72.17 | 88.77 | 92.85 | 64.33 | 62.64 | 79.53 | 85.32 | 36.50 | 60.00 | 82.10 | 87.90 | 47.37 |
| CSNR (Ours) | **74.71** | 90.12 | **94.16** | 66.52 | **66.74** | **81.81** | **86.88** | 40.06 | **65.00** | **84.75** | **90.60** | **50.51** |

# 4 Experiments

## 4.1 Experiments Setting

We evaluate our method on the CUHK-PEDES [16], ICFG-PEDES [4], and RSTPReid [35] datasets. In terms of dataset partition as well as data augmentation, our approach follows the previous work [9]. To ensure a fair comparison, the method is benchmarked against other approaches that also employ CLIP as the pre-trained model. For evaluation, we adopt Rank-1, Rank-5, Rank-10, and mean Average Precision (mAP) to comprehensively assess retrieval performance. The parameters mentioned earlier are set as follows: $M$ =0.40, $\mu = 0.25$, $\eta = 0.30$, $\alpha = 0.15$, $K = 0.50$, $\tau = 0.02$.

Table 2: Ablation study results on the ICFG-PEDES dataset.

| No. | CF | FF | CSSM | NSRM | Loss | R1 | R5 | R10 | mAP |
|---|---|---|---|---|---|---|---|---|---|
| 1 | ✓ | × | × | × | WKCL | 63.25 | 79.34 | 84.53 | 36.26 |
| 2 | ✓ | ✓ | × | × | WKCL | 65.77 | 81.06 | 85.99 | 38.82 |
| 3 | ✓ | ✓ | × | × | Triplet loss | 65.62 | 81.13 | 85.56 | 38.24 |
| 4 | ✓ | ✓ | × | × | E-LOSS | 65.39 | 81.14 | 85.82 | 38.59 |
| 5 | ✓ | ✓ | ✓ | × | WKCL | 66.61 | 81.53 | 86.57 | 39.48 |
| 6 | ✓ | ✓ | ✓ | ✓ | WKCL | 66.74 | 81.81 | 86.88 | 40.06 |

Table 3: Analysis of average image-text similarity across different sample types.

| Dataset | Clean Sample | Noisy Sample | New Clean Sample |
|---|---|---|---|
| CUHK-PEDES | 0.5181 | 0.3958 | 0.5666 |
| ICFG-PEDES | 0.5782 | 0.4253 | 0.5971 |
| RSTPReid | 0.4922 | 0.3123 | 0.5306 |


Figure 4: Visualization results before and after the NSRM. The rematched descriptions are more appropriate to the original descriptions.

## 4.2 Comparison with State-of-the-art Methods

We present the performance of our proposed method alongside other state-of-the-art models in Table 1. CFine [28] leverages the full potential of CLIP's knowledge, while IRRA [9] utilizes fine-grained interaction to enhance global alignment. TBPS-CLIP [2] conducts empirical studies, such as data augmentation, to improve the model's performance. The best result is highlighted in bold.

Notably, our method achieves superior performance compared to existing approaches published in the last two years. On the RSTPReid dataset, our model surpasses the second-best method by significant margins of +3.05% Rank-1, +1.20% Rank-5, +0.89% Rank-10, and +2.25% mAP, respectively. This is because RSTPReid contains a greater number of noisy samples compared to other datasets, which poses a greater challenge and thus allows our method to demonstrate a more significant improvement. These consistent improvements validate the effectiveness of addressing the noisy descriptions problem in TBPR.

## 4.3 Ablation Study

As reported in Table 2, an ablation study is performed on the ICFG-PEDES dataset. We can draw the following observations: (1) It is evident from the comparison of No. 1 and No. 2 that using both CF and FF results in better performance. (2) By comparing No. 2 and No. 5, we can see that CSSM improves the model's performance. Specifically, CSSM mitigates the negative impact of the noisy samples. (3) Analyzing the results from No. 5 and No. 6, it is apparent that NSRM improves the model's performance by effectively converting some noisy samples into clean ones, as illustrated in Figure 4. (4) By comparing No. 2, No. 3, and No. 4, we observe that WKCL outperforms Triplet loss and E-Loss. The reason is that WKCL not only leverages more information on negative samples but also prioritizes the harder negatives to ensure the efficiency of model training under noisy description circumstances.

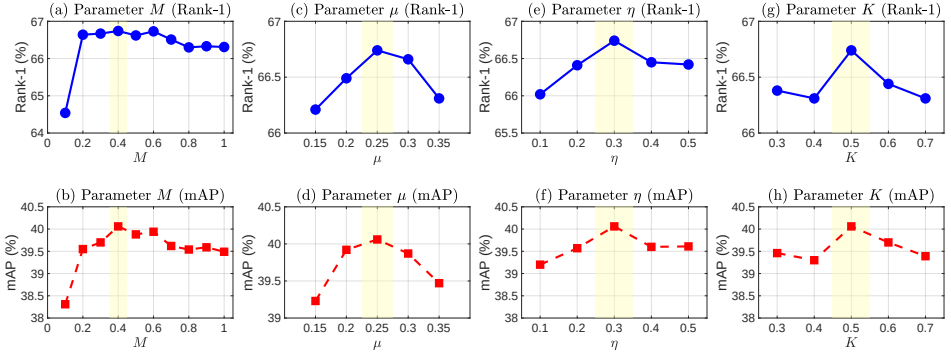


Figure 5: Parameter analysis on ICFG-PEDES dataset.

## 4.4 Parametric Analysis

There are some important hyperparameters in the proposed method, *i.e.*, the $M$ in feature extractor, the $\mu$ in CSSM, the $\eta$ in NSRM, and the $K$ in WKCL. We can draw the following conclusions from Figure 5: (1) When $M$ is too small, the feature extractor will lose too many fine-grained tokens. Otherwise, redundant fine-grained tokens impair the feature quality. (2) When $\mu$ is set too low, many clean samples may be incorrectly classified as noisy samples. Conversely, a significant number of noisy samples might be mistakenly identified as clean samples. (3) Small $\eta$ values weaken the NSRM's impact. When $\eta$ is too large, the image-text similarity of the newly generated samples may even become worse than that of the original samples, failing to meet the first transformation condition of NSRM mentioned earlier. So the NSRM stops generating new samples. That is why, when $\eta$ exceeds 0.4, the decline becomes less significant, as shown in (e) and (f). (4) On the one hand, if $K$ is too small, WKCL fails to utilize sufficient negative sample information. On the other hand, if $K$ is too large, it introduces too many non-hard negative samples into training, which can degrade model performance.

## 4.5 Further Analysis

To further validate the effectiveness of CSSM and NSRM, we measured the average cosine similarity between images and texts for: (1) clean samples identified by CSSM, (2) noisy samples filtered by CSSM, and (3) newly generated samples from NSRM across three datasets. The evaluation was conducted using data from the best-performing epoch. As shown in Table 3, the clean samples selected by the CSSM demonstrate significantly higher image-text cosine similarity compared to the noisy samples. Moreover, the NSRM-generated clean samples achieve even higher similarity scores than those of CSSM-identified clean samples. Given the notably poor quality of the RSTPReid dataset, it can be observed that the image-text similarity of clean samples is consistently and significantly higher than that of noisy samples. These noisy samples, which tend to severely degrade model performance, highlight the robustness of our approach. This further explains why our method achieves such remarkable improvement on the RSTPReid benchmark. These results provide additional evidence supporting the efficacy of both CSSM and NSRM.

# 5 Conclusion

In the field of TBPR, the presence of noisy samples can significantly degrade model performance, but this issue is often overlooked. To address this problem, we propose a robust framework called CSNR. The CSSM filters out noisy samples to ensure the quality of the training data. The NSRM converts some noisy samples into clean samples, thereby increasing the set of clean samples. The WKCL leverages more negative information and pays more weight to harder negatives, making model training more efficient, under circumstances with noisy descriptions. Experimental results confirm the effectiveness of our method.

# Acknowledgement

This work was supported in part by the Sichuan Science and Technology Program under Grant 2024ZYD0263, in part by the Open Foundation of Yunnan Key Laboratory of Software Engineering under Grant 2023SE206.

# References

[1] Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, and Simon Lacoste-Julien. A closer look at memorization in deep networks. In *Proc. 34th Int. Conf. Mach. Learn.*, page 233–242, 2017.

[2] Min Cao, Yang Bai, Ziyin Zeng, Mang Ye, and Min Zhang. An empirical study of clip for text-based person search. In *Proc. AAAI Conf. Artif. Intell.*, pages 465–473, 2024.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proc. Conf. North Am. ChapterAssoc. Comput. Linguist.: Hum. Lang. Technol.*, pages 4171–4186, 2019.

[4] Zefeng Ding, Changxing Ding, Zhiyin Shao, and Dacheng Tao. Semantically self-aligned network for text-to-image part-aware person re-identification. *arXiv preprint arXiv:2107.12666*, 2021.

[5] Changsheng Gao, Yiheng Jiang, Siqi Wu, Yifan Ma, Li Li, and Dong Liu. IMOFC: Identity-level metric optimized feature compression for identification tasks. *IEEE Trans. Circuits Syst. Video Technol.*, pages 1–1, 2024.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 770–778, 2016.

[7] Shuting He, Hao Luo, Wei Jiang, Xudong Jiang, and Henghui Ding. VGSG: Vision-guided semantic-group network for text-based person search. *IEEE Transactions on Image Processing*, 33:163–176, 2024.

[8] Zhenyu Huang, Guocheng Niu, Xiao Liu, Wenbiao Ding, Xinyan Xiao, Hua Wu, and Xi Peng. Learning with noisy correspondence for cross-modal matching. *Adv. Neural Inf. Process. Syst.*, pages 29406–29419, 2021.

[9] Ding Jiang and Mang Ye. Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pages 2787–2797, 2023.

[10] Min Jiang, Xinyu Liu, and Jun Kong. Text-guided prototype generation for occluded person re-identification. *IEEE Signal Process. Lett.*, 2024.

[11] Sejun Kim, Sungjae Kang, Hyomin Choi, Seong Soo Kim, and Kisung Seo. Keypoint aware robust representation for transformer-based re-identification of occluded person. *IEEE Signal Process. Lett.*, pages 65–69, 2023.

[12] Jiayi Li, Min Jiang, Jun Kong, Xuefeng Tao, and Xi Luo. Learning semantic polymorphic mapping for text-based person retrieval. *IEEE Trans. Multimedia*, pages 10678–10691, 2024.

[13] Junnan Li, Richard Socher, and Steven C. H. Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. In *Proc. Int. Conf. Learn. Represent.*, 2020.

[14] Shangze Li, Andong Lu, Yan Huang, Chenglong Li, and Liang Wang. Joint token and feature alignment framework for text-based person search. *IEEE Signal Process. Lett.*, pages 2238–2242, 2022.

[15] Shuang Li, Tong Xiao, Hongsheng Li, Wei Yang, and Xiaogang Wang. Identity-aware textual-visual matching with latent co-attention. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 1908–1917, 2017.

[16] Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang. Person search with natural language description. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 1970–1979, 2017.

[17] Yating Liu, Zimo Liu, Xiangyuan Lan, Wenming Yang, Yaowei Li, and Qingmin Liao. Dm-adapter: Domain-aware mixture-of-adapters for text-based person retrieval. In *Proc. AAAI Conf. Artif. Intell.*, pages 5703–5711, 2025.

[18] Yu Liu, Guihe Qin, Haipeng Chen, Zhiyong Cheng, and Xun Yang. Causality-inspired invariant representation learning for text-based person retrieval. In *Proc. AAAI Conf. Artif. Intell.*, volume 38, pages 14052–14060, 2024.

[19] Haim Permuter, Joseph Francos, and Ian Jermyn. A study of gaussian mixture models of color and texture features for image classification and segmentation. *Pattern Recognit. Lett.*, pages 695–706, 2006.

[20] Yang Qin, Yingke Chen, Dezhong Peng, Xi Peng, Joey Tianyi Zhou, and Peng Hu. Noisy-correspondence learning for text-to-image person re-identification. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pages 27187–27196, 2024.

[21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proc. Int. Conf. Mach. Learn.*, pages 8748–8763, 2021.

[22] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 815–823, 2015.

[23] Mengyang Sun, Wei Suo, Peng Wang, Kai Niu, Le Liu, Guosheng Lin, Yanning Zhang, and Qi Wu. An adaptive correlation filtering method for text-based person search. *Int. J. Comput. Vis.*, pages 1–16, 2024.

[24] Di Wang, Feng Yan, Yifeng Wang, Lin Zhao, Xiao Liang, Haodi Zhong, and Ronghua Zhang. Fine-grained semantics-aware representation learning for text-based person retrieval. In *Proc. Int. Conf. Multimedia Retrieval*, page 92–100, 2024.

[25] Fangyu Wu, Jeremy S. Smith, Wenjin Lu, Chaoyi Pang, and Bailing Zhang. Attentive prototype few-shot learning with capsule network-based embedding. In *Proc. Eur. Conf. Comput. Vis.*, page 237–253, 2020.

[26] Xinyi Wu, Wentao Ma, Dan Guo, Tongqing Zhou, Shan Zhao, and Zhiping Cai. Text-based occluded person re-identification via multi-granularity contrastive consistency learning. In *Proc. AAAI Conf. Artif. Intell.*, pages 6162–6170, 2024.

[27] Chao Xie and Tomoki Toda. Noisy-to-noisy voice conversion under variations of noisy condition. *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, pages 3871–3882, 2023.

[28] Shuanglin Yan, Neng Dong, Liyan Zhang, and Jinhui Tang. CLIP-Driven fine-grained text-image person re-identification. *IEEE Trans. Image Process.*, 32:6032–6046, 2023.

[29] Shuanglin Yan, Jun Liu, Neng Dong, Liyan Zhang, and Jinhui Tang. Prototypical prompting for text-to-image person re-identification. In *Proc. ACM Int. Conf. Multimedia*, pages 2331–2340, 2024.

[30] Shuanglin Yan, Neng Dong, Shuang Li, and Huafeng Li. Trimatch: Triple matching for text-to-image person re-identification. *IEEE Signal Process. Lett.*, pages 1–5, 2025.

[31] Shuo Yang, Zhaopan Xu, Kai Wang, Yang You, Hongxun Yao, Tongliang Liu, and Min Xu. BiCro: Noisy correspondence rectification for multi-modality data via bi-directional cross-modal similarity consistency. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pages 19883–19892, 2023.

[32] Kaiyang You, Wenjing Chen, Chenji Wang, Hao Sun, and Wei Xie. Cross-modal feature fusion-based knowledge transfer for text-based person search. *IEEE Signal Process. Lett.*, 2024.

[33] Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor Tsang, and Masashi Sugiyama. How does disagreement help generalization against label corruption? In *Proc. Int. Conf. Mach. Learn.*, pages 7164–7173, 2019.

[34] Xu Zhang, Hao Li, and Mang Ye. Negative pre-aware for noisy cross-modal matching. In *Proc. AAAI Conf. Artif. Intell.*, pages 7341–7349, 2024.

[35] Aichun Zhu, Zijie Wang, Yifeng Li, Xili Wan, Jing Jin, Tian Wang, Fangqiang Hu, and Gang Hua. DSSL: Deep surroundings-person separation learning for text-based person retrieval. In *Proc. ACM Int. Conf. Multimedia*, pages 209–217, 2021.

[36] Jialong Zuo, Hanyu Zhou, Ying Nie, Feng Zhang, Tianyu Guo, Nong Sang, Yunhe Wang, and Changxin Gao. UFineBench: Towards text-based person retrieval with ultra-fine granularity. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pages 22010–22019, 2024.