

MINING FALSE POSITIVE EXAMPLES FOR TEXT-BASED PERSON RE-IDENTIFICATION

Wenhao Xu, Zhiyin Shao, Changxing Ding

South China University of Technology
School of Electronic and Information Engineering
Guangzhou, Guangdong, China

ABSTRACT

Text-based person re-identification (ReID) aims to identify images of the targeted person from a large-scale person image database according to a given textual description. However, due to significant inter-modal gaps, text-based person ReID remains a challenging problem. Most existing methods generally rely heavily on the similarity contributed by matched word-region pairs, while neglecting mismatched word-region pairs which may play a decisive role. Accordingly, we propose to mine false positive examples (MFPE) via a jointly optimized multi-branch architecture to handle this problem. MFPE contains three branches including a false positive mining (FPM) branch to highlight the role of mismatched word-region pairs. Besides, MFPE delicately designs a cross-reLU loss to increase the gap of similarity scores between matched and mismatched word-region pairs. Extensive experiments on CUHK-PEDES demonstrate the superior effectiveness of MFPE. Our code is released at <https://github.com/xx-adeline/MFPE>.

Index Terms— Text-based Person Retrieval, Person Re-identification, multi-granularity image-text alignments

1. INTRODUCE

Text-based person re-identification (ReID) is aimed at identifying the targeted person images from a large-scale person image database according to a given textual description. It is a powerful video surveillance tool and has drawn increasing attention from both academia and industry recently.

Unfortunately, text-based person ReID is still a challenging problem due to fine-grained problem. To be specific, it is difficult to distinguish between people who are dressed very similarly, such as the two people shown in Fig. 1. They wear black shirts, black pants, and white shoes as described in the text query. The only difference is that the man on the left is holding a red umbrella and the man on the right is holding a jacket. In this case, most existing methods generally ignore a problem [1], [2], [3], [4], [5], [6], [7]. The matched word-region pairs get high similarity scores and contribute a lot to the final instance-level similarity, while the mismatched word-region pairs have little effect on it. This means that all

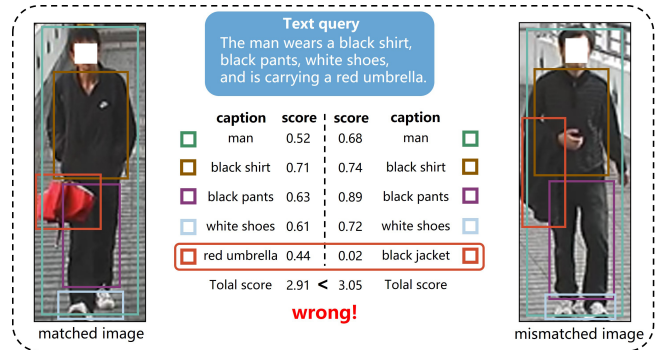


Fig. 1. The two people are difficult to distinguish based on the text query. The person on the left has the same identity as the text, while the person on the right has a different identity.

mismatched word-region pairs are probably to be completely ignored. However, a matched image-text pair should not have any mismatched word-region pairs. The ignored word-region pairs are inevitably prone to cause false-positive matching. Therefore, it is essential to highlight the role of mismatched word-region pairs to downgrade the overall similarity of mismatched image-text pairs.

To this end, we propose to mine false positive examples (MFPE) via a jointly optimized multi-branch architecture for text-based person ReID. As illustrated in Fig. 2, MFPE contains three branches. The global and local branches are for aligning visual features and text features and calculating the similarity of image-text pairs [1]. In the false positive mining (FPM) branch, we delicately design a novel cross-reLU loss to forcedly delineate a clear decision boundary and maximally increase the gap of similarity scores between matched and mismatched word-region pairs which are sampled in a balanced manner. Based on the discriminative similarity, we can precisely mine mismatched word-region pairs and utilize them as a bias to modify the overall similarity. To demonstrate the efficacy of MFPE, we conduct extensive experiments on the CUHK-PEDES database. The results show that MFPE can effectively mine false positive examples and outperforms compared methods.

2. RELATED WORKS

Text-based person ReID was first introduced by Li et al. [8]. They proposed a GNA-RNN model to calculate the affinity between each image-text pair and collected a large-scale person description dataset called CUHK-PEDES. Later, to alleviate the fine-grained problem caused by all samples belonging to a single category, the region-word-based method and the region-phrase-based method are popular for text-based person ReID. For example, Niu et al. [2] proposed a Multi-granularity Image-text Alignments (MIA) model to align cross-modal features at three different granularities. Wang et al. [9] aligned body regions with noun phrases with the help of a light auxiliary attribute segmentation layer and a natural language parser. SSAN [1] was proposed to automatically extract region-level textual features for its corresponding visual regions by introducing extra prediction to the word-region correspondences. Zhu et al. [4] proposed a DSSL model to explicitly separate surroundings information and person information to obtain higher retrieval accuracy. Recently, Wang et al.

Recently, Zhang et al. [10] proposed a negative-aware attention framework. However, we found through experiments that as a model for image-text retrieval, its data sampling method and attention mechanism are not suitable for text-based ReID. For example, NAAF judges whether the word-region matches based on the principle that all word-region pairs of a mismatched image-text pair mismatch. It is obviously inappropriate for person ReID because all samples belong to the same category, and regions generally have matched words. Inspired by this, we explore mining false positive examples for text-based ReID.

3. METHOD

3.1. Global and local branches

Visual Representation Extraction: We utilize a pretrained ResNet-50 [11] to extract visual feature maps $F \in \mathbb{R}^{H \times W \times C}$. To obtain the global visual representation $V_g \in \mathbb{R}^P$, we first perform Global Max Pooling (GMP) to downsample on F . And then we reduce it to P -dim through a 1×1 conv layer. For the local branch [1], we first horizontally partition F into K non-overlapping regions $V_e = \{v_k\}_{k=1}^K$ [12] and separately embed them through K corresponding 1×1 conv layers.

Textual Representation Extraction: We utilize a Bi-LSTM [13] to extract the sentence representation $E \in \mathbb{R}^{C \times n}$ after the words are embedded via a pretrained BERT language model [14]. Similar to the global visual branch, we perform Row-wise Max Pooling (RMP) and a 1×1 conv layer to obtain the global textual representation $T_g \in \mathbb{R}^P$. For the local branch, we adopt a Word Attention Module (WAM) [1] to modify E to K local textual representations. We also use

RMP and K corresponding 1×1 conv layers on modified representations to get final local textual representations.

When it comes to calculating similarity, we respectively concatenate the K regions visual and textual representations V_l and $T_l \in \mathbb{R}^{K \times P}$ into the unified representations. Eventually, the global and local similarity of an image-text pair can be calculated as:

$$s_g = \frac{V_g^T T_g}{\|V_g\| \|T_g\|}, \quad s_l = \frac{V_l^T T_l}{\|V_l\| \|T_l\|}. \quad (1)$$

3.2. FPM branch

As discussed in the Introduction, the goal of the FPM branch is to highlight the effect of mismatched word-region pairs in an effective way. The first step is to mine the mismatched word-region pairs through the similarity scores. Concretely, we separately project local visual representations $V_e = \{v_k\}_{k=1}^K$ and textual representations $E = \{e_i\}_{i=1}^n$ into a common feature space via a 1×1 conv layer and then compute the semantic similarity scores as:

$$s_{k,i} = \frac{\theta(v_k)^T \phi(e_i)}{\|\theta(v_k)\| \|\phi(e_i)\|}, \quad k \in [1, K], i \in [1, n], \quad (2)$$

where $\theta(v_k) = W_\theta v_k$, $\phi(e_i) = W_\phi e_i$. $W_\theta, W_\phi \in \mathbb{R}^{M \times C}$.

For text-based ReID, the number of regions is so limited that the region that mismatches any word is almost non-existent. Therefore, we mine mismatched word-region pairs by searching for words that mismatch any region. we perform a max-pooling operation on similarity scores because the maximum similarity between the word and all regions is low, indicating that the word mismatches any region.

$$s_i = \max_k(s_{k,i}). \quad (3)$$

The experimental analysis in section 4.3 demonstrates that a learnable decision boundary is unnecessary, so zero is considered as the decision boundary for matched and mismatched word-region pairs. In order to enhance the discriminative power of mismatched pairs, the **mining mask** operation is employed to filter out negative similarities. Eventually, we use them to calculate the instance-level similarity and modify the local similarity:

$$s_{neg} = \sum_{i=1}^n Mask_{mining}(s_i), \quad s_{local-neg} = s_l + s_{neg}, \quad (4)$$

where $Mask_{mining}(\cdot)$ denotes that when the input is positive, it is 0, and when the input is negative, it is unchanged.

During inference, the overall similarity of an image-text pair is the sum of s_g , s_l , and $s_{local-neg}$.

3.3. Optimization

Instead of simply utilizing the similarity scores to distinguish whether the word-region pairs match, we propose a

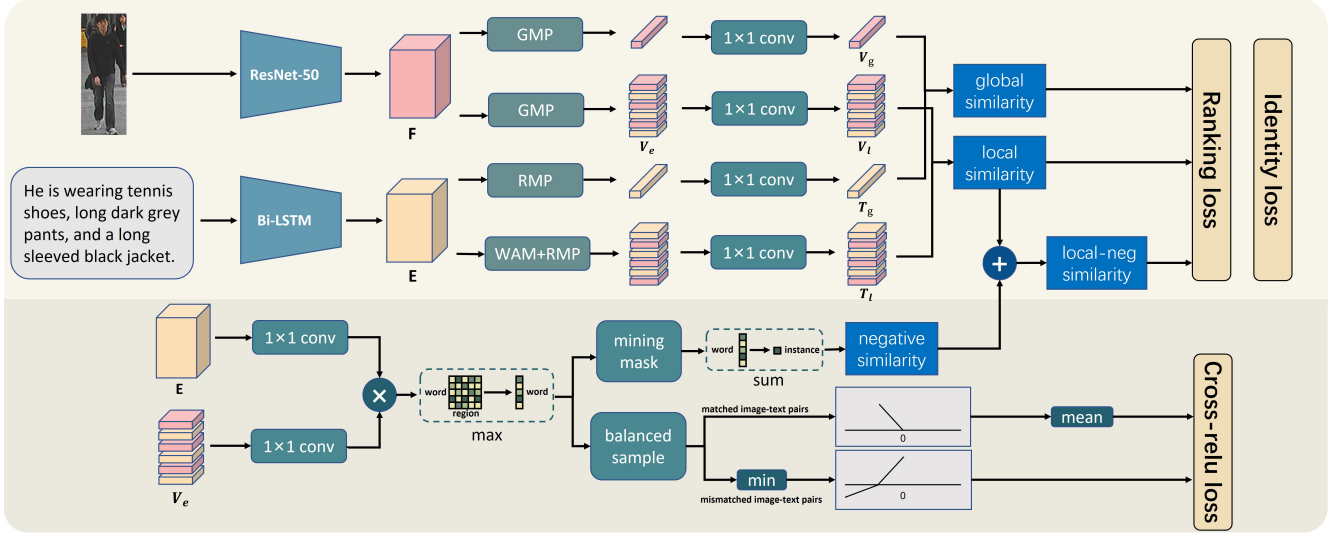


Fig. 2. The overall framework of our proposed MFPE model, containing a global branch, a local branch, and an FPM branch to jointly infer the similarity of an image-text pair.

novel **cross-relu loss** to increase the similarity gap between them. Specifically, We first adopt a **balanced sample** strategy to obtain the same number of matched and mismatched image-text pairs in a mini-batch. Since there are no labels for word-region pairs, the coarse judgment of whether word-region pairs match is based on the following principles: 1) All word-region pairs of a matched image-text pair match. 2) Mismatched image-text pair has at least one mismatched word-region pair.

For the matched image-text pairs, we adopt a relu with slope m_1 and bias b_1 to train their similarity scores tending to be positive:

$$L_m = \frac{1}{n} \sum_{i=1}^n \max(-m_1 s_i + b_1, 0), \quad (5)$$

where slope m_1 and bias b_1 are empirically set to 1 and 0.001 (The function is shown in Fig. 2) to ensure that the similarity scores keep away from the sensitive decision boundary.

For the mismatched image-text pairs, we first perform a min operation to obtain the most probable mismatched word-region pair. Similarly, we train it to be negative via a leaky relu with slope m_2 and bias b_2 :

$$s_{min} = \min_i(s_i), \quad L_{mm} = \max(m_2 s_{min} + b_2, 0), \quad (6)$$

where slope m_2 and bias b_2 are empirically set to 1 and 0.15. Note that adopting leaky relu instead of relu and setting b_2 large enough is aimed at maximizing the mining of mismatched word-region pairs at the cost of the accuracy of matched pairs.

To further optimize MFPE, the popular identity loss is employed in global and local branches.

$$L_{id}(x) = -\log(\text{softmax}(W_{id}x)), \quad (7)$$

$$L_{id} = L_{id}(V_g) + L_{id}(T_g) + \lambda_1(L_{id}(V_l) + L_{id}(T_l)), \quad (8)$$

where W_{id} is a shared FC layer between the two modalities and λ_1 is set to 0.5.

Besides, the popular ranking loss is adopted to constraint the intra-class similarity score to be larger than the inter-class similarity with a margin α :

$$L_r(s) = \max(\alpha - s(V_+, T_+) + s(V_+, T_-), 0) + \max(\alpha - s(V_+, T_+) + s(V_-, T_+), 0), \quad (9)$$

$$L_r = L_r(s_g) + \lambda_2 L_r(s_l) + \lambda_3 L_r(s_{local-neg}), \quad (10)$$

where (V_+, T_-) and (V_-, T_+) denote a mismatched image-text pair while (V_+, T_+) denotes a matched image-text pair. $S(\cdot, \cdot)$ stands for the similarity of a pair. Note that λ_2 and λ_3 are set to 0.5 and 0.25 to avoid $s_{local-neg}$ replacing s_l .

4. EXPERIMENTS

4.1. Experiment Settings

Dataset and Evaluation Metrics: We conduct extensive experiments on CUHK-PEDES [8] dataset and adopt the popular Recall at K ($R@K$, $K=1, 5, 10$) to evaluate performance. Following the official evaluation protocol, the training set contains 34,054 images and 68,126 textual descriptions for 11003 persons. The validation and test sets include data for 1,000 persons, respectively.

Implementation Detail: All experiments are conducted on an NVIDIA Titan Xp GPU. We adopt Adam as the optimizer with the initial learning rate of 0.001. The mini-batch size and number of epochs are empirically set to 64 and 45. Following previous methods [1], C , K , P , n , α , and M are set to 2048, 6, 1024, 100, 0.2, and 256, respectively.

4.2. Comparison Results

We compare our proposed MFPE method with previous approaches on the CUHK-PEDES database in Table 1. Comparison results show MFPE outperforms all other methods. For example, compared with the typical region-based model MIA, MFPE obtains a significant 10.72% improvement on R@1. Moreover, compared with SSAN, MFPE still achieves nearly 2.45% improvement in terms of R@1.

4.3. Ablation Study

In the following, we conduct extensive ablation studies on CUHK-PEDES to analyze the effectiveness of each branch in Table 2. It is obvious that the local similarity modified by negative similarity achieves a significant improvement of 1.40% on R@1 compared with the unmodified. The ‘global + local + FPM’ also promotes the R@1 accuracy of the ‘global + local’ by 2.70%. The above improvement strongly demonstrates the effectiveness of our proposed framework for mining false positive examples.

Moreover, we conduct a series of ablation studies to analyze the effectiveness of each component of the FPM branch in Table 3. 1) The baseline model refers to MFPE without the FPM branch and BERT in the training stage. 2) When removing the ranking loss on local-negative similarity, robustness and performance are severely degraded since the modified local similarity lacks a direct constraint. 3) Mining mask is a key component in the FPM branch to emphasize the effect of mismatched pairs, without which the performance drops by 1.34% 4) Without balanced sampling, the performance is slightly decreased, due to the apparent propensity for mismatched data. But the cost of time increases by about 4 times. 5) Adding a learnable decision boundary to MFPE will obtain suboptimal performance. However, The boundary is always kept around the initial value of zero. Consequently, the learnable decision boundary is unnecessary.

Table 1. Performance Comparisons on CUHK-PEDES.

Methods	ref	R@1	R@5	R@10
GNA-RNN [15]	CVPR17	19.05	-	53.64
CMPM/C [16]	ECCV18	49.37	71.69	79.27
MIA [2]	TIP20	53.10	75.00	82.90
SCAN [17]	ECCV18	55.86	75.97	83.69
SUM [18]	KBS22	59.22	80.35	87.60
DSSL [4]	MM21	59.98	80.41	87.56
MGEL [19]	IJCAI21	60.27	80.01	86.74
SSAN [1]	arXiv21	61.37	80.15	86.73
LapsCore [20]	ICCV21	63.40	-	87.80
MFPE w/o BERT(ours)	-	61.92	80.80	87.25
MFPE(ours)	-	63.82	82.63	88.66

Table 2. Ablation study about branches on CUHK-PEDES.

Global	Local	FPM	bert	R@1	R@5	R@10
✓	-	-	-	54.68	75.42	82.73
-	✓	-	-	57.57	77.06	84.86
✓	✓	-	-	59.22	78.72	85.65
-	✓	✓	-	58.97	78.44	85.49
✓	✓	✓	-	61.92	80.80	87.25
✓	✓	✓	✓	63.82	82.63	88.66

Table 3. Ablation study about the FPM branch design on CUHK-PEDES.

Methods	R@1	R@5	R@10
Baseline	59.22	78.72	85.65
w/o local-negative ranking loss	59.81	79.22	86.24
w/o mining mask	60.58	80.77	87.17
w/o balanced sample	61.22	80.30	87.28
w learnable decision boundary	61.47	80.70	87.35
Full	61.92	80.80	87.25

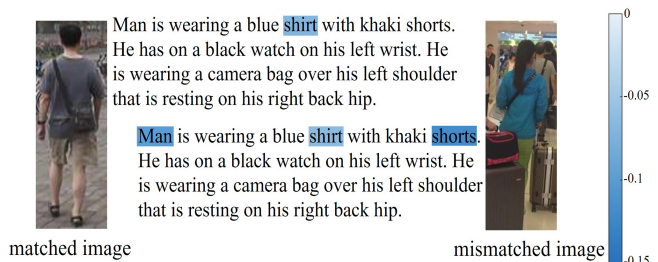


Fig. 3. Visualization of the word-region negative similarity of a matched image-text pair and a mismatched image-text pair.

4.4. Visualization

We visualize the word-region negative similarity according to a given text query in Fig. 3. For the matched image, almost all words obtain non-negative similarity. As for the mismatched image, obviously mismatched words like ‘MAN’ and ‘short’ obtain a corresponding negative similarity. This case demonstrates the effectiveness of the FPM branch for mining mismatched pairs.

5. CONCLUSION

In this paper, we propose to mine false positive examples (MFPE) via a jointly optimized multi-branch architecture for text-based person ReID. Specifically, MFPE employs global and local branches to extract semantically aligned features and delicately designs an FPM branch to mine and emphasize the effect of mismatched word-region pairs. Moreover, We introduce a novel cross-reLU loss to increase the gap of similarity scores between matched and mismatched word-region pairs. Finally, we conduct extensive experiments to demonstrate the effectiveness of MFPE.

6. REFERENCES

- [1] 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.
- [2] Kai Niu, Yan Huang, Wanli Ouyang, and Liang Wang, “Improving description-based person re-identification by multi-granularity image-text alignments,” *IEEE Transactions on Image Processing*, vol. 29, pp. 5542–5556, 2020.
- [3] Surbhi Aggarwal, Venkatesh Babu Radhakrishnan, and Anirban Chakraborty, “Text-based person search via attribute-aided matching,” in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2020, pp. 2617–2625.
- [4] 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 *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 209–217.
- [5] Chenyang Gao, Guanyu Cai, Xinyang Jiang, Feng Zheng, Jun Zhang, Yifei Gong, Pai Peng, Xiaowei Guo, and Xing Sun, “Contextual non-local alignment over full-scale representation for text-based person search,” *arXiv preprint arXiv:2101.03036*, 2021.
- [6] Kecheng Zheng, Wu Liu, Jiawei Liu, Zheng-Jun Zha, and Tao Mei, “Hierarchical gumbel attention network for text-based person search,” in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 3441–3449.
- [7] Kai Niu, Yan Huang, and Liang Wang, “Textual dependency embedding for person search by language,” in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 4032–4040.
- [8] Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang, “Person search with natural language description,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.
- [9] Zhe Wang, Zhiyuan Fang, Jun Wang, and Yezhou Yang, “Vitaa: Visual-textual attributes alignment in person search by natural language,” in *European Conference on Computer Vision*. Springer, 2020, pp. 402–420.
- [10] Kun Zhang, Zhendong Mao, Quan Wang, and Yongdong Zhang, “Negative-aware attention framework for image-text matching,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 15661–15670.
- [11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [12] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang, “Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline),” in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 480–496.
- [13] Sepp Hochreiter and Jürgen Schmidhuber, “Long short-term memory,” *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.
- [15] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, Mingliang Xu, and Yi-Dong Shen, “Dual-path convolutional image-text embeddings with instance loss,” *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 16, no. 2, pp. 1–23, 2020.
- [16] Ying Zhang and Huchuan Lu, “Deep cross-modal projection learning for image-text matching,” in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 686–701.
- [17] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He, “Stacked cross attention for image-text matching,” in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 201–216.
- [18] Zijie Wang, Aichun Zhu, Jingyi Xue, Daihong Jiang, Chao Liu, Yifeng Li, and Fangqiang Hu, “Sum: Serialized updating and matching for text-based person retrieval,” *Knowledge-Based Systems*, vol. 248, pp. 108891, 2022.
- [19] Chengji Wang, Zhiming Luo, Yaojin Lin, and Shaozi Li, “Text-based person search via multi-granularity embedding learning,” in *IJCAI*, 2021, pp. 1068–1074.
- [20] Yushuang Wu, Zizheng Yan, Xiaoguang Han, Guanbin Li, Changqing Zou, and Shuguang Cui, “Lapscore: language-guided person search via color reasoning,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1624–1633.