Part-Based Multi-Scale Attention Network for Text-Based Person Search

Yubin Wang, Ding Qi, and Cairong Zhao(✉)

Department of Computer Science and Technology, Tongji University, Shanghai, China
{1851731,2011267,zhaocairong}@tongji.edu.cn

Abstract. Text-based person search aims to retrieve the target person in an image gallery based on textual descriptions. Solving such a fine-grained cross-modal retrieval problem is very challenging due to differences between modalities. Moreover, the inter-class variance of both person images and descriptions is small, and more semantic information is needed to assist in aligning visual and textual representations at different scales. In this paper, we propose a **P**art-based **M**ulti-Scale **A**ttention **N**etwork (PMAN) capable of extracting visual semantic features from different scales and matching them with textual features. We initially extract visual and textual features using ResNet and BERT, respectively. Multi-scale visual semantics is then acquired based on local feature maps of different scales. Our proposed method learns representations for both modalities simultaneously based mainly on Bottleneck Transformer with self-attention mechanism. A multi-scale cross-modal matching strategy is introduced to narrow the gap between modalities from multiple scales. Extensive experimental results show that our method outperforms the state-of-the-art methods on CUHK-PEDES datasets.

Keywords: Person re-identification · Cross-modal retrieval · Representation learning

1 Introduction

Recently, text-based person search has gained increasing attention due to its potential applications in intelligent surveillance. It aims to retrieve the target person according a relevant textual description. Since natural language is more accessible as retrieval queries, text-based person search has great necessity in the absence of target images. However, it is a challenging task due to difficulties of both person re-identification and cross-modal retrieval. First, it is difficult to extract robust features due to interference from occlusion and background clutter. Second, all images and descriptions belong to the same category, person, thus making inter-modality variance much larger than intra-modality variance.

To solve these problems, related methods [1–8] have been proposed in recent years to reduce the gap between these two modalities and thus improve the

S. Yu et al. (Eds.): PRCV 2022, LNCS 13534, pp. 462–474, 2022.
https://doi.org/10.1007/978-3-031-18907-4_36

matching accuracy. These methods always focus on two problems, one is how to learn representations in a coarse-to-fine manner for both modalities, and the other is how to find an adaptive multi-scale cross-modal matching strategy that all features are well-aligned. Many current works are unable to solve these two problems well at the same time. Some of them learn representations only from local scale [1–3] or global scale [4–7], which are unable to generate features at different scales from both coarse-grained and fine-grained perspectives. Although some approaches [9,10] consider combining local features with global features, some fragments of textual descriptions still cannot align with visual regions that are semantically consistent with them.

The relevance at different scales makes it difficult to align visual and textual features. For multi-scale matching, existing methods [2–5] try to align images and texts at different scales using predefined rules. However, these methods do not take into account the cross-scale association between modalities. As shown in Fig. 1, images and textual descriptions can be decomposed into regions and phrases at local scale. Since the phrase "belt" exists in one visual region while the phrases "long sleeve white shirt" and "black pants" appear in two separated visual regions, phrase-region matching at a fixed scale is not effective, where the cross-scale matching of semantics between modalities is completely ignored.

Retaining semantics for visual representation learning is always critical. Some methods [11,12] use horizontal segmentation referring to PCB [13] in person re-identification, aiming to match relevant textual semantics based on local salient parts of images. However, this segmentation operation can easily break visual semantics existing in different regions. As shown in Fig. 1, the visual semantics of the phrase "books" is exactly partitioned by two visual regions, and from neither of these two regions can the model accurately recognize the semantics matching the phrase. This prevents the model from fully extracting key information, and leads to inaccurate matching results. Considering the above, an approach is urgently needed for multi-scale feature extraction while preserving semantics at different scales.

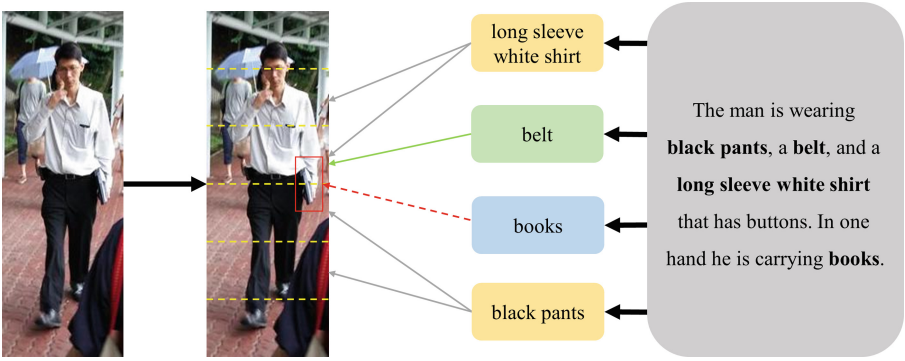

Fig. 1. Phrase-region matching at a fixed scale is not effective, where the cross-scale semantic matching between modalities is completely ignored, and the segmentation operation can easily break visual semantics existing in different regions.

To address these problems, we introduce a part-based multi-scale attention network for text-based person search, aiming at improving the representation learning and matching methods from different scales in an end-to-end manner and enhancing the semantics with the ability to collect global information by self-attention mechanism [14]. For visual representation learning, we use a pre-trained ResNet [15] to generate the basic feature map for each image, which is horizontally segmented into several strips to generate regions at different scales. All visual features with the same scale are combined together after scale-specific attention-based branches with Bottleneck Transformer [16] blocks to output visual representations. For textual representation learning, each word embedding is learned by a pre-trained BERT [17] with fixed parameters and is further processed by a network with hybrid branches. In each branch, textual representations adaptively learn to match visual representations, thus eliminating the inter-modality variance. In addition, we introduce a multi-scale cross-modal matching strategy with the cross-modal projection matching (CMPM) loss [4], thus gradually reducing the gap between modalities from different scales. Our main contributions are summarized as follows.

- We propose a dual-path feature extraction framework for learning multi-scale visual and textual representations simultaneously, where semantic information is captured for both modalities based on Bottleneck Transformer blocks with self-attention mechanism.
- We introduce a multi-scale cross-modal matching strategy using cross-modal projection matching (CMPM) loss, thus gradually reducing the variance between modalities from different scales.
- Our proposed method outperforms all other methods on the CUHK-PEDES [6] datasets. Extensive ablation studies demonstrate the effectiveness of components in our method.

2 Related Works

2.1 Person Re-identification

Recently, there are many person re-identification methods based on deep learning to improve the matching accuracy by exploring and mining the fine-grained discriminative features in person images. PCB [13] proposes a convolutional baseline based on local information, which segments the global feature map into horizontal strips to extract local features. MGN [18] varies the number of divided strips in different branches to obtain local feature representations with multiple scales. In addition, some works [19–21] consider the detection of body parts with external tools or attention mechanism to improve the quality of local features by detecting subtle changes in local regions. However, such approaches rely heavily on pose estimation and semantic parsing algorithms, while ignoring the semantic connection between different local regions, resulting in critical visual semantics not being fully extracted. Moreover, many works [13,22] tend to limit to a fixed local scale without paying attention to the semantic information at other scales, thus reducing the discrimination of representations.

2.2 Text-Based Person Search

The development of text-based person search is gradually gaining attention from the research community. Li et al. [6] first introduce the text-based person search task and propose GNA-RNN to output the similarity between images and textual descriptions. Zheng et al. [5] propose a dual-path convolutional neural network for visual-linguistic embedding learning, which can be efficiently fine-tuned end-to-end. Zhang et al. [4] design cross-modal projection matching (CMPM) loss and cross-modal projection classification (CMPC) loss for cross-modal embedding learning. Some works are based on body parts with external tools to assist in extracting visual features. Among them, PMA [2] proposes a pose-guided multi-granularity attention network to match visual regions associated with descriptions from multiple granularities based on human pose estimation. VITAA [3] uses semantic segmentation labels to drive the learning of attribute-aware features.

Some recent works have focused more on feature matching at different scales. AXM-Net [9] dynamically exploits multi-scale knowledge from both modalities and recalibrates each modality based on shared semantics. NAFS [23] constructs full-scale representations for visual and textual representations and adaptively conducts joint alignments at all scales. SSAN [10] extracts semantic alignment by exploring relatively aligned body parts as supervision and using contextual cues from descriptions to extract part features. TIPCB [12] learns visual and textual local representations through a dual-path local alignment network structure with a multi-stage cross-modal matching strategy. Nevertheless, such methods lack attention to semantic integrity, leading to inaccurate alignment. By comparison, we propose a novel method that learns and aligns representations more effective by utilising self-attention mechanism within multi-scale setting.

3 Our Approach

In this section, we explain our PMAN in detail. First, we introduce the framework for extracting visual and textual representations. Then we describe the multi-scale cross-modal matching module. The architecture is shown in Fig. 2.

3.1 Multi-scale Visual Representation Learning

For visual representation learning, we first take the image I as the input of ResNet, and the feature $f_I \in R^{H \times W \times C}$ generated after its fourth residual blocks is used as the basic feature map, where H, W and C represent the dimension of height, width and channel. The structure of the multi-scale visual representation learning module based on this feature map, including a local-scale branch, a medium-scale branch and a global-scale branch, is shown in Fig. 3.

For each visual branch, we utilise Bottleneck Transformer [16] (BoT) as backbone, which introduces the self-attention mechanism into the bottleneck architecture by replacing the convolutional layer with a multi-headed self-attention

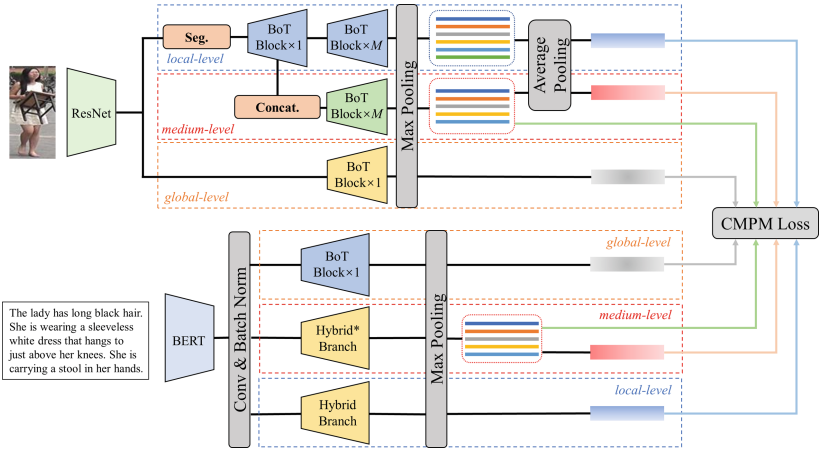

Fig. 2. The architecture of the proposed PMAN, including a dual-path feature extraction framework with attention-based branches for multiple scales and a multi-scale cross-modal matching module. BoT stands for Bottleneck Transformer. **Seg.** indicates horizontal segmentation on the basic feature map. **Concat.** indicates concatenation of every two neighboring local feature maps. * indicates parallel branches.

layer (MHSA). The CNN backbone, due to the nature of the convolutional kernel, tends to focus on local features instead of semantic integrity, so it is essential to stack Transformer blocks that are specialize in capturing global information within a specific scale, thus achieving better performance with less parameters.

In the local-scale branch, We first use the strategy of PCB [13] to horizontally segment the basic feature map f_I into several local regions $\{f_{I,i}\}_{i=1}^{K}$, where $f_{I,i} \in R^{\frac{H}{K} \times W \times C}$. Then we take each local region as the input of a Bottleneck Transformer consisting of $M+1$ blocks to obtain local features $\{f_{I,i}^{l}\}_{i=1}^{K}$ after a maximum pooling layer, where $f_{I,i}^{l} \in R^{1 \times C}$. These features usually contain fine-grained semantics and play a crucial role in learning discriminative visual features.

In the medium-scale branch, considering that the semantics existing in multiple local regions are easily destroyed, we concatenate every two neighboring local feature maps after the first block of the local-scale branch to generate medium-scale regions. We take them as the input of a Bottleneck Transformer consisting of M blocks to obtain self-attention weighted feature maps for medium-scale regions, and output $K-1$ medium-scale features $\{f_{I,i}^{m}\}_{i=1}^{K-1}$ after a maximum pooling layer, where $f_{I,i}^{m} \in R^{1 \times C}$, which usually contain the significant semantic information associated with descriptions. By combining semantic information in local-scale regions, the semantics disrupted by segmentation is preserved. For above two branches, local-scale representation $f_I^l \in R^C$ and medium-scale representation $f_I^m \in R^C$ are generated after an average pooling layer.

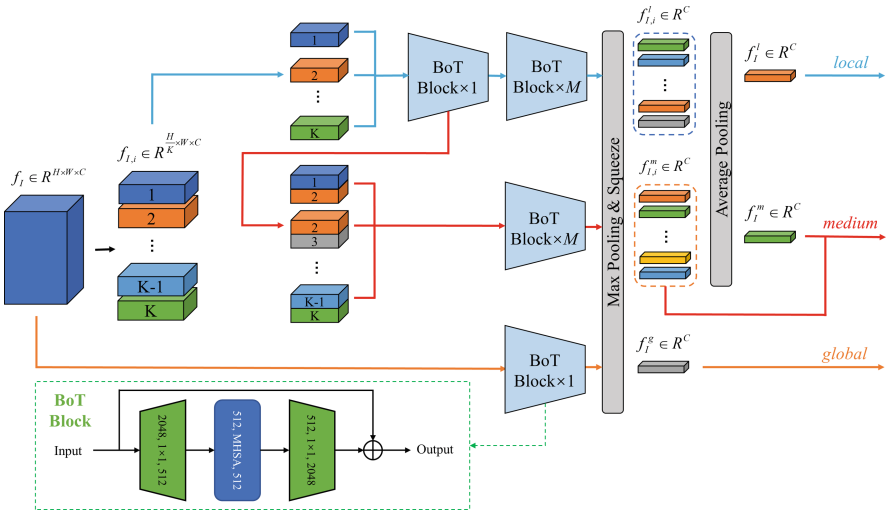

Fig. 3. The structure of the multi-scale visual feature learning module. BoT stands for Bottleneck Transformer. MHSA stands for multi-head self-attention.

In the global-scale branch, we directly use the basic feature map f_I as the input to the Bottleneck Transformer. Since this feature map already contains semantic information at global level, the Transformer for this branch consists of one single block to highlight the visual semantic information. The global-scale feature is obtained after a maximum pooling layer, and squeezed to global-scale representation $f_I^g \in R^C$, which is not influenced by local semantics. These above three representations as well as the set of medium-scale features serve for the multi-scale matching stage.

3.2 Multi-scale Textual Representation Learning

For texual representation learning, a pre-trained language model BERT [17] is used to extract word embeddings with discriminative properties. Specifically, we decompose the sentence and tokenize it to obtain the token sequence, and then truncate or pad the sequence according to the maximum length L. We feed them into BERT with fixed parameters to generate word embeddings $f_w \in R^{L \times D}$, where D denotes the dimension of each word embedding. Since the dimension of embeddings needs to match with the input of bottleneck blocks, we expand the word embeddings and pass them through an 1×1 convolutional network to adjust the dimension of channel from D to C, and textual feature $f_t \in R^{1 \times L \times C}$ is obtained after a batch normalization layer.

Similar to visual representation learning, our multi-scale textual representation learning module also consists of local-scale branch, medium-scale branch and global-scale branch, each for adapting visual representations at the same scale. We refer to the method introduced by Chen et al. [18], which stacks residual bottlenecks as textual backbone, in which way can textual representations adaptively learn to match visual representations. We improve this approach by introducing a novel hybrid branch, as shown in Fig. 4. The hybrid branch consists of two residual bottlenecks and a Bottleneck Transformer block sandwiched between them. For efficiency, the former residual bottleneck shares the parameter with all branches at the same scale. Considering the effectiveness for visual recognition, we apply Bottleneck Transformer to textual representation learning, aiming to extract long-distance relations between word embeddings for better comprehension. By mixing the residual bottleneck and Bottleneck Transformer, we introduce self-attention into local feature learning in a multi-scale way.

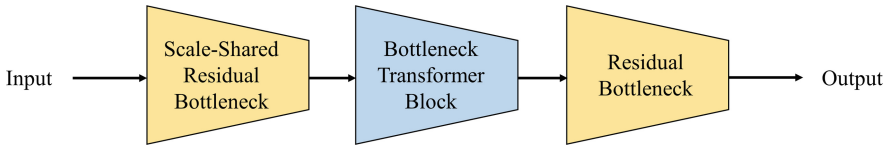

Fig. 4. The structure of the hybrid branch, consisting of a scale-shared residual bottleneck, a Bottleneck Transformer block and a residual bottleneck.

The local-scale and medium-scale branches consist of one single hybrid branch and $K-1$ paralleled hybrid branches respectively, and a maximum pooling layer is added at the end of each branch to generate textual representations $f_t^l \in R^C$ and $\{f_{t,i}^m\}_{i=1}^{K-1}$, where $f_{t,i}^m \in R^C$. Medium-scale textual representations are further processed with an average pooling layer to generate representation $f_t^l \in R^C$ for matching. In the global-scale branch, considering that stacking complex blocks tends to be time-consumed for training while accuracy is not significantly improved, we only use one single Bottleneck Transformer block for learning long-distance association between word embeddings, while suppressing the overfitting phenomenon. A maximum pooling layer is then added to obtain the global-scale representation $f_t^g \in R^C$. The representations extracted from these three branches are already capable of adapting to the visual representations, so that textual representations with highly integrated visual semantics can be learned.

3.3 Multi-scale Feature Matching

In the multi-scale feature matching stage, we use the cross-modal projection matching (CMPM) loss [4] as the loss function, which minimizes the KL divergence between the projection compatibility distributions and the normalized matching distributions to eliminate the difference between textual and visual modalities. Specifically, within a given small batch of N pairs, according to the

image representation x_i^I, the set of image-text representation pairs within the batch can be denoted as $\left\{\left(x_i^I, x_j^T\right), y_{i,j}\right\}_{j=1}^N$, where $y_{i,j} = 1$ when x_i^I and x_j^T have the same identity, and $y_{i,j} = 0$ otherwise. For each image-text representation pair, we can calculate the matching probability between them by:

$$p_{i,j} = \frac{\exp\left(x_i^{I\top} \bar{x}_j^T\right)}{\sum_{k=1}^N \exp\left(x_i^{I\top} \bar{x}_k^T\right)} \quad s.t. \quad \bar{x}_j^T = \frac{x_j^T}{\left\|x_j^T\right\|}, \tag{1}$$

where $\bar{x}_j^T$ denotes the regularized textual representation. For the matching probability between x_i^I and x_j^T, we use the normalized label distribution $q_{i,j}$ as the real distribution. The matching loss in one direction can be calculated by:

$$L_{i2t} = \sum_{i=1}^N \sum_{j=1}^N p_{i,j} \log \frac{p_{i,j}}{q_{i,j} + \varepsilon} \quad s.t. \quad q_{i,j} = \frac{y_{i,j}}{\sum_{k=1}^N y_{i,k}}, \tag{2}$$

where ε is a very small value to avoid numerical problems. L_{t2i} can be calculated in a reverse way. Therefore, the CMPM loss is computed by $L = L_{i2t} + L_{t2i}$.

Considering that our dual-path feature extraction framework consists of several branches, each of which generates features at specific scale, we sum the CMPM loss according to visual and textual representations from all scales. The overall objective function is calculated by:

$$L_{overall} = L_g + L_m + L_l + L_{align}, \tag{3}$$

where L_g, L_m and L_l indicate the CMPM loss of local-scale, medium-scale and global-scale representations, respectively. The aligning loss $L_{align} = \sum_{i=1}^{K-1} L_{m,i}$ represents the CMPM loss computed with medium-scale features, which preserve the semantic association between local regions and have more information aligned with descriptions comparing to other scales. By reducing the overall function, our model learns well-aligned representations for both modalities.

4 Experiments

4.1 Experimental Setup

Datasets and Evaluation Protocol. The CUHK-PEDES [6] datasets, which is currently the mainstream benchmark for text-based person search, contains 40206 images of 13003 person IDs, each of which has two textual descriptions annotated by different annotators. These textual descriptions have a vocabulary of 9408 different words. The training set has 34054 images of 11003 person IDs. The validation set has 3078 images of 1000 person IDs, and the test set has 3074 images of 1000 person IDs.

The experimental results are measured by the top-K (K = 1, 5, 10) metric. Given a textual description as query, all test images are ranked according to their similarity to the query, and top-K indicates the percentage of successful searches among all searches in the first K results.

Implementation Details. For visual representation learning, the input images are resized to 384 × 128. We use ResNet50 [15] as visual backbone to extract the basic feature map. The height, width and channel dimension of the basic feature map are set to $H = 24$, $W = 8$, $C = 2048$. The number of local-scale visual regions is set to $K = 6$. The number of Bottleneck Transformer blocks is set to $M = 3$. For textual representation learning, we use a pre-trained BERT-Base-Uncase for extracting word embeddings, where the maximum length is set to $L = 64$. In the training phase, we use an SGD optimizer with momentum to optimize the model for 80 epochs. The initial learning rate is 0.003 decreased by 0.1 after 50 epochs. We randomly horizontally flip and crop the images to augment data. In the testing phase, we simply sum the local-scale, medium-scale and global-scale representations as the final representation for retrieval. The batch size for both the training and testing phase are set to $N = 64$.

4.2 Comparison with State-of-the-Art Methods on CUHK-PEDES

This section compares the result of our method proposed in this paper with other previous works, as shown in Table 1. These methods can be broadly classified into global-level methods and local-level methods. Compared with global-level methods, local-level methods prefer to obtain discriminative features from local visual regions and align them with phrase-level semantics in textual descriptions to improve matching accuracy. It can be observed that our PMAN can outperform all the existing methods. This further illustrates that our multi-scale approach with self-attention is crucial for improving matching accuracy.

Table 1. Comparison with state-of-the-art methods on CUHK-PEDES datasets. Top-1, top-5 and top-10 accuracies (%) are reported. "g" represents the methods only using global features, and "l+g" represents the methods using global and local features.

Method	Type	Top-1	Top-5	Top-10
GNA-RNN [6]	g	19.05	–	53.64
IATV [7]	g	25.94	–	60.48
PWM + ATH [24]	g	27.14	49.45	61.02
Dual Path [5]	g	44.40	66.26	75.07
CMPM + CMPC [4]	g	49.37	–	79.27
MIA [1]	l + g	53.10	75.00	82.90
PMA [2]	l + g	53.81	73.54	81.23
ViTAA [3]	l + g	55.97	75.84	83.52
NAFS [23]	l + g	59.94	79.86	86.70
MGEL [11]	l + g	60.27	80.01	86.74
SSAN [10]	l + g	61.37	80.15	86.73
AXM-Net [9]	l + g	61.90	79.41	85.75
LapsCore [25]	l + g	63.40	–	87.80
TIPCB [12]	l + g	64.26	**83.19**	89.10
PMAN (Ours)	l + g	**64.51**	83.14	**89.15**

4.3 Ablation Studies

Effects of Multi-scale Representation Learning. In this section, we conduct experiments for each branch to analyze the importance of representation learning at different scales. Table 2 shows the experimental results on the CUHK-PEDES dataset when different branches are selected to train for representation learning. From the variants (a), (b) and (c), it shows that when only one single branch is trained, the medium-scale branch gains the best accuracy, which proves that it is crucial to preserve the semantic association between local regions. From the variants (b) and (g), it can be seen that the multi-scale feature learning framework can enhance the discrimination of visual and textual features at different scales, thus enabling our model to fuse intra-modality features to improve the matching accuracy while aligning inter-modality semantic information.

Table 2. Performance comparison of training with different branches in our method. Top-1, top-5 and top-10 accuracies (%) are reported.

Variant	Local-scale	Medium-scale	Global-scale	Top-1	Top-5	Top-10
(a)	✓			60.12	81.36	88.13
(b)		✓		62.26	82.33	88.59
(c)			✓	60.55	81.34	88.26
(d)	✓	✓		62.07	82.26	88.52
(e)	✓		✓	61.79	82.27	88.55
(f)		✓	✓	63.10	82.71	88.96
(g)	✓	✓	✓	**64.51**	**83.14**	**89.15**

Effects of Backbone with Self-attention. To demonstrate the importance of self-attention mechanism for extracting semantic representations, we analyze the impact of Bottleneck Transformer structure employed as backbone for our proposed model. For comparison, we replace Bottleneck Transformer blocks with residual bottleneck blocks in all branches of different modalities, while other components remain unchanged. The experimental results are shown in Table 3. Comparing the variant (a) with (b) and (c), it can be found that Bottleneck Transformer blocks with self-attention in both modalities improve the matching accuracy, which proves that this design enables the association of salient semantics between modalities. Moreover, the self-attention in textual representation learning has a greater impact on results, which reveals the fact that the structure can extract long-distance relations between word embeddings, which make them produce higher response when relevant queries are given. From the variant (d), it can be seen that the attention-based architecture can facilitate our model to extract better semantic information from both modalities, thus achieving an excellent matching accuracy in the multi-scale representation matching stage.

Table 3. Performance comparison of learning visual and textual representations with Bottleneck Transformer as backbone in our method. Top-1, top-5 and top-10 accuracies (%) are reported.

Variant	Visual backbone	Textual backbone	Top-1	Top-5	Top-10
(a)			61.63	81.65	88.30
(b)	✓		62.31	82.35	88.73
(c)		✓	62.68	82.42	88.94
(d)	✓	✓	**64.51**	**83.14**	**89.15**

5 Conclusion

In this paper, we propose a part-based multi-scale attention network capable of extracting visual semantic features from different scales and matching them with textual features. For representation learning, we introduce Bottleneck Transformer with self-attention mechanism in both modalities to capture features with semantics. For representation matching, we adopt a multi-scale cross-modal adaptive matching strategy. The comparison results show that our approach outperforms the state-of-the-art methods on CUHK-PEDES dataset. Extensive ablation studies demonstrate the effectiveness of components in our method.

References

1. Niu, K., Huang, Y., Ouyang, W., Wang, L.: Improving description-based person re-identification by multi-granularity image-text alignments. IEEE Trans. Image Process. **29**, 5542–5556 (2020)
2. Jing, Y., Si, C., Wang, J., Wang, W., Wang, L., Tan, T.: Pose-guided multi-granularity attention network for text-based person search. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 11189–11196 (2020)
3. Wang, Z., Fang, Z., Wang, J., Yang, Y.: *ViTAA*: visual-textual attributes alignment in person search by natural language. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12357, pp. 402–420. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58610-2_24
4. Zhang, Y., Lu, H.: Deep cross-modal projection learning for image-text matching. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11205, pp. 707–723. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01246-5_42
5. Zheng, Z., Zheng, L., Garrett, M., Yang, Y., Xu, M., Shen, Y.D.: Dual-path convolutional image-text embeddings with instance loss. ACM Trans. Multimedia Comput. Commun. Appl. (TOMM) **16**(2), 1–23 (2020)
6. Li, S., Xiao, T., Li, H., Zhou, B., Yue, D., Wang, X.: Person search with natural language description. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1970–1979 (2017)
7. Li, S., Xiao, T., Li, H., Yang, W., Wang, X.: Identity-aware textual-visual matching with latent co-attention. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1890–1899 (2017)

8. Aggarwal, S., Radhakrishnan, V.B., Chakraborty, A.: Text-based person search via attribute-aided matching. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 2617–2625 (2020)
9. Farooq, A., Awais, M., Kittler, J., Khalid, S.S.: AXM-Net: cross-modal context sharing attention network for person Re-ID. arXiv preprint arXiv:2101.08238 (2021)
10. Ding, Z., Ding, C., Shao, Z., Tao, D.: Semantically self-aligned network for text-to-image part-aware person re-identification. arXiv preprint arXiv:2107.12666 (2021)
11. Wang, C., Luo, Z., Lin, Y., Li, S.: Text-based person search via multi-granularity embedding learning. In: IJCAI (2021)
12. Chen, Y., Zhang, G., Lu, Y., Wang, Z., Zheng, Y.: TIPCB: a simple but effective part-based convolutional baseline for text-based person search. Neurocomputing (2022)
13. Sun, Y., Zheng, L., Yang, Y., Tian, Q., Wang, S.: Beyond part models: person retrieval with refined part pooling (and a strong convolutional baseline). In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11208, pp. 501–518. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01225-0_30
14. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems 30 (2017)
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
16. Srinivas, A., Lin, T.Y., Parmar, N., Shlens, J., Abbeel, P., Vaswani, A.: Bottleneck transformers for visual recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16519–16529 (2021)
17. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
18. Wang, G., Yuan, Y., Chen, X., Li, J., Zhou, X.: Learning discriminative features with multiple granularities for person re-identification. In: Proceedings of the 26th ACM International Conference on Multimedia, pp. 274–282 (2018)
19. Zhao, H., et al.: Spindle net: person re-identification with human body region guided feature decomposition and fusion. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1077–1085 (2017)
20. Song, G., Leng, B., Liu, Y., Hetang, C., Cai, S.: Region-based quality estimation network for large-scale person re-identification. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32 (2018)
21. Kalayeh, M.M., Basaran, E., Gökmen, M., Kamasak, M.E., Shah, M.: Human semantic parsing for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1062–1071 (2018)
22. Fu, Y., et al.: Horizontal pyramid matching for person re-identification. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 8295–8302 (2019)
23. Gao, C., et al.: Contextual non-local alignment over full-scale representation for text-based person search. arXiv preprint arXiv:2101.03036 (2021)

24. Chen, T., Xu, C., Luo, J.: Improving text-based person search by spatial matching and adaptive threshold. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1879–1887. IEEE (2018)
25. Wu, Y., Yan, Z., Han, X., Li, G., Zou, C., Cui, S.: LapsCore: language-guided person search via color reasoning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1624–1633 (2021)