Attentive Feature Focusing for Person Search by Natural Language

Renjie Pan[1,2], Hua Yang[1,2(✉)], Xinxin Yang[1,2], Mao Xiaodong[3], Long Ye[4], and Da Pan[4]

[1] The Institute of Image Communication and Network Engineering, Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai, China
{rjpan21,hyang,yang-xinxin}@sjtu.edu.cn

[2] Shanghai Key Laboratory of Digital Media Processing and Transmission, Shanghai Jiao Tong University, Shanghai, China

[3] Science and Technology Department of Shanghai Public Security Bureau, Shanghai, China

[4] State Key Laboratory of Media Convergence and Communication, Communication University of China, Beijing 100024, China
{yelong,pdmeng}@cuc.edu.cn

Abstract. Cross-modal Retrieval has made great improvement since social media and artificial intelligence is gradually changing our life style. Human interaction is no longer limited to a single mode. Cross-modal retrieval brings greater portability to users. Extracting more sufficient feature from multimedia data and better fusing the feature have become the focus of academic research. In response to these two issues, this paper applies deep learning as the foundation for person search by natural language, combining CNN and Transformer. Meanwhile, we utilize BERT, the most effective feature extraction network in Natural Language Processing and proposed a network architecture using attention mechanism. The framework is divided into a global branch and a fine-grained branch. At the same time, a human parsing network is designed to classify different parts of the characters to obtain feature representation in attribute level. In the visual-textual alignment, the k-reciprocal sampling algorithm is used to construct a more comprehensive sample pair. The proposed framework has achieved *state-of-the-art* performance in CUHK-PEDES, which proves its great capacity and prospect.

Keywords: Person search by natural language · Cross-modal retrieval · Metric learning

This work was supported in part by National Natural Science Foundation of China (NSFC, Grant No. 61771303, 62171281), Science and Technology Commission of Shanghai Municipality (STCSM, Grant No. 19DZ1209303, 20DZ1200203, 2021SHZDZX0102), SJTU Yitu/Thinkforce Joint Laboratory for Visual Computing and Application, and the Open Research Project of the State Key Laboratory of Media Convergence and Communication, Communication University of China, China (No. MCCSKLMCC2020KF003).

G. Zhai et al. (Eds.): IFTC 2021, CCIS 1560, pp. 266–280, 2022.
https://doi.org/10.1007/978-981-19-2266-4_21

1 Introduction

The goal of cross-modal retrieval is to use data from one modality to retrieve another one. The major challenge is measuring the similarity between two modalities. Therefore, a more comprehensive modeling is indispensable during cross-modal retrieval. The current research goal is to design more effective ways to make it more accurate and expandable. Traditional retrieval tasks can only be based on a single modality. For example, in pedestrian re-identification(Re-ID), police officers can only describe the appearance of the suspect based on the verbal description of the witness. In this case, cross-modal retrieval becomes significant in real life.

Whether the model can merge the feature from multi modalities is directly related to the retrieval performance [23]. Taking text-image retrieval as an example, images of different identities may have human characteristics such as hairstyle, facial expression, skin color, tops, bottoms, shoes, bags, etc. Descriptions of images include color, shape, posture, etc. Thanks to deep learning and computer vision, the feature extraction backbone learns abundant feature from images and text such as ResNet [7] and LSTM [8], which solves gradient disappearance generated by RNN [18] and better deal with long sequences. The rich time-domain feature is widely used in the industry. In the past two years, transformer [21] has made a breakthrough in temporal feature representation. In cross-modal retrieval, better fusing two modalities is the main direction of current research.

There are still many challenge in cross-modal retrieval. For example, if a single samples in two modalities share with similar characteristics, chances are that they may be mismatched. This situation is called "mal-positioned matching". In Visual Question Answering (VQA) [3] and image-text retrieval [10], the overall similarity across modalities is measured by the vectors of the image and textual features. The negative sample that is very similar to the positive sample is false positive pair, which makes it difficult for the model to distinguish effectively, thus cannot match them with specific appearance attributes. This problem has made researchers quite confused in recent years, which poses a nontrivial mission for fine-grained interaction between images and text.

This work analyzes the characteristics of existing cross-modal retrieval algorithms and launches an in-depth study of person search by natural language algorithms and proposes a framework using the most advanced backbones, which optimize the text feature extraction capabilities and hence better integrate textual features with image features and improve the comprehensive performance. The main contributions of the work can be summarized as follows:

- **Transformer combined with CNN.** In the image backbone, we fuse CNN with Transformer, which combines both the ability of processing local feature representation and the ability of processing long-term feature representation.
- **BERT as textual feature backbone.** The textual feature backbone uses BERT [5] as the textual feature backbone network, which utilizes its larger vocabulary corpus, pays better attention to the characteristics of long-

sequence, captures more precise semantic information and effectively improves the overall performance.

- **Global-grained branch combined with fine-grained branch.** The local attribute obtained by the Human-parsing network (HRNet) serves as a segmentation block to facilitate each fine-grained branch to create a segmentation map of a specific attribute category. It serves as a knowledge regulator to present the attribute of fine-grained branch. At the same time, contrastive learning is also used in fine-grained branch in order to obtain good embedding of both images and text and to learn excellent feature representations in conjunction with global-grained branch.

2 Related Works

2.1 Cross-Modal Retrieval

Cross-modal retrieval aims at fusing the data between two modalities to accomplish the retrieval task of a given modality and another one. There are plenty of multimedia data such as images, text, voices, videos, physiological signals, etc. Figure 1 shows the cross-modal retrieval framework under common circumstances.

Cross-modal retrieval of images and text requires that both images and text have good feature representation. Traditional image backbone [13] generates local feature representation by detecting key points in the images. Many methods are proposed to handle correlative tasks, which can be generally concluded into two categories: 1) Associative subspace learning [12,24,25,27] and 2) Coupled similarity learning [11,14,19].

The purpose of associative subspace learning is to find an associative latent space where the embedding of images and text can be directly compared. It usually measures the distance based on the characteristics of the two modalities. Andrew et al. proposed Deep Canonical Correlation Analysis (DCCA) [1], which learns complex nonlinear projections of different forms of data to obtain linearly correlated representations. In addition, [27] further extended [1]. But DCCA causes instability of the covariance of each batch will cause eigenvalue problem. Bi-directional ranking loss [12,24,25] optimizes triplet loss [19]. Wang et al. [25] designed a dual-channel nonlinear neural network to extract features of text and images separately, with bi-directional ranking loss as the objective function. Later, [24] used embedded and similarity network to learn a latent semantic embedding, in which the similarity network learned the similarity through the dot product of textual and visual feature matrices. Liu et al. [12] designed the Recurrent Residual Fusion (RRF) module based on the original dual-channel neural network architecture. It draws on residual module idea in ResNet and adds RRF to the network, which makes a better representation of images and text and greatly innovates the previous work.

Coupled similarity learning aims to design a similarity network that can predict image-text matching pairs. In addition to measuring the global-grained similarity between two modalities, many research works are devoted to better

matching local regions and textual fragments. Ma et al. [14] introduced CNN for the first time in the image-text retrieval matching problem, proposed m-CNN and segmented the textual content to match the image features. However, it did not pay enough attention to visual fine-grained semantics.

Li et al. [11] compares the feature of a mini-batch with all N samples in another modality. Based on this, CMCE also designed a two-stage network: 1) CMCE loss function, which was designed to learn the identity-aware feature representation of images and texts to obtain ideal text and image feature representations. 2) Further capture fine-grained alignments, which focus more on word-level similarity. The second stage is tightly coupled by CNN-LSTM in order to match remarkable regions and latent text semantics. In addition, the second stage pays more attention to negative samples, so only hard negative samples can enter the second stage for learning, which is a good solution to the negative sample selection strategy. However, it requires hardware of high level and computing power requirements to store samples. Similarly, [19] proposed category prediction to learn the image-text embedding layer. Huang et al. proposed sm-LSTM [14], which includes a multimodal context-modulated attention scheme at each time step. By predicting saliency maps, an image-text pair can be selectively focused on. The representation of the selected pair is obtained based on the predicted saliency map, and then they are compared to measure the local similarity. sm-LSTM focused on the potential information in text through innovative textual feature backbone networks and image instance perception network.

2.2 Person Search by Natural Language

The current cross-modal retrieval between images and text can be categorized into two forms, which are searching by images and natural language. The former can also be termed as Person re-identification (Re-ID).

The typical Re-ID methods refer to the retrieval of the candidates with the highest correlation in the image datasets. However, clear and effective images cannot always be obtained easily, which hinders the application and development of Re-ID. Recently, researchers have focused their attention on re-recognition through text descriptions: recognizing pedestrians through natural language descriptions. At the same time, the model also faces huge challenges such as processing more complicated long and random sentences, and low-quality surveillance images may lead to inconsistent interpretation. In order to solve these problem, [9,12,17] proposed attention mechanism to build a modular connection between visual and textual representation, [15] proposed a cross-modal objective function, and [4] adopted tailoring in order to discover dense visual features, and learns the matching mode at the regional level [30]. In addition, [28] also introduced pose estimation for more details of human body.

It is very important to use appropriate feature representations for image-text retrieval. Many studies have consistently adopted the global image features and only use word segmentation on text description. As a result, some visual cues are ignored in the expression of the fine-grained level and thus visual-textual feature expression is not comprehensive.

For textual feature representation, the phrases serve as the corresponding component of the specific visual cues and is usually the ground truth. Another method is to extract textual features by segmenting the sentence first and then identifying the nominal phrases. Textual attributes are treated as auxiliary tags to supplement the content of image features.

We know that the information we can tell from the text must be greater than or equal to the information expressed by the image. Otherwise, image-text pairs must contain mismatched information and can only be processed as a negative sample. Therefore, whether textual feature is fully discovered completely affects the matching performance. Moreover, there is a fundamental difference between text descriptions even if they look very similar, which brings more difficulty in distinguishing them. Long and random sentences are tend to learn redundant information, which is an urgent problem.

3 Proposed Method

The overall structure of the proposed model is shown in Fig. 1. In global-grained branch, we extract the entire image and the whole text to obtain the global feature representation. Specifically, given an image of a pedestrian, we use ResNet with pre-trained parameters as the basic network. On this basis, we add a Transformer layer, which act as an attention mechanism. Eventually, the visual representation is mapped into an embedding space through a fully connected layer, and the global visual representation $x_g \in R^2$ is obtained. Given a textual description, we segment it first and embed the words into BERT to capture the contextual relations of each words. The hidden states of the antecedent and subsequent word are linked. We obtain the intermediate textual features of all connected hidden states through max-pooling. Then, by mapping the textual representation into the embedding space through a MLP layer, a global-grained textual representation $z_g \in R^{512}$ is generated.

In fine-grained branch, we use Human-parsing network (HRNet) to classify the characters according to different parts of the pedestrians and obtain the fine-grained visual representation; At the same time, the whole sentence are segmented and divided into words. Words are classified into attribute level, which obtains the fine-grained representation of a single word.

3.1 Word Segmentation

Given a natural language description, we use Pytorch transformers to segment the text first. We use BERT as the textual backbone to fully extract textual feature representation. Then, we use Stanford POS tagger to parse and extract the nominal phrases in each text and obtain the global textual information and the fine-grained textual information at the same time.

Then we use a clustering dictionary to extract words with particular information or original attributes in a single sentence. Specifically, we collect a word list of each attribute category, such as “jeans”, “sports shoes”, etc. We use word

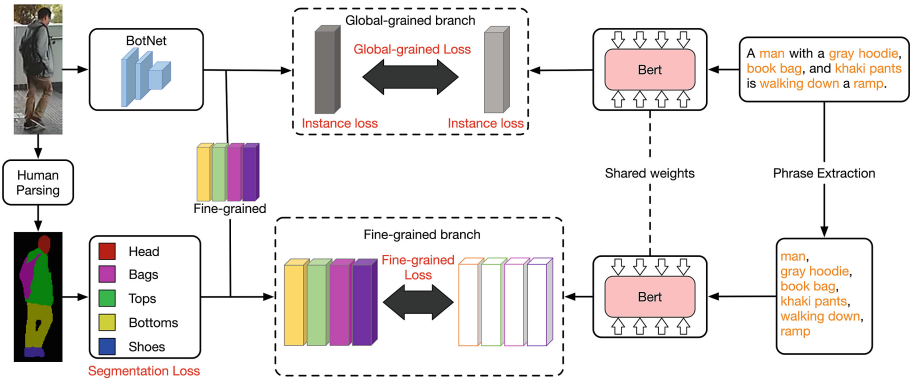

Fig. 1. The proposed architecture is composed of visual and textual input. The image-text embedding learns the joint feature representation and visual-textual alignment module consists of both global and fine-grained branches.

vectors of the same attribute category as anchor embedding. Based on this, the attributes in each category are highly related words and phrases. The extracted attributes are conducive to matching with the classification of human body parts so as to fully achieve a fine-grained analysis. Using clustering to process words at attribute level is one of the innovations of this research.

3.2 Attention Mechanism

Attention mechanism has recently achieved great success in deep neural networks. It was first proposed in machine translation, namely sequence to sequence(seq2seq) tasks. seq2seq treats the last output of the encoder as the input of the decoder, compresses the entire content of the original sequence into a fixed-size vector, which cannot fully mine the information of long sequences. Attention mechanism [2] is one of the core innovations of machine translation. Attention mechanism alleviates this by allowing the decoder to look back at the hidden state of the original sequence, and then provide weight average as an additional input to the decoder, making full use of the output features of each hidden layer in the decoding stage.

Based on Attention mechanism, Transformer also uses an encoder-decoder architecture, but it is more complicated. Whose specific model is shown in Fig. 2. Multi-head self-attention enables current node not only pay attention to the current word, but also obtain the semantics of the context. It uses different linear transformations to project Query, Key and Value, and finally stitches different results together, details as follows:

$$MH(Q, K, V) = Con(h_1, h_2, \ldots, h_h)W^0 \tag{1}$$

$$h_i = Att(QW_i^Q, KW_i^K, VW_i^V) \tag{2}$$

where h_i stands for each head, Con for concatenate, MH for Multi head algorithm and Att for Attention module. At the same time, each sub-network uses residual links, so the sub-network output can be then denoted as:

$$sub_{layeroutput} = LayerNorm(x + (SubLayer(x))) \tag{3}$$

The structure of decoder resembles the encoder, in addition that a self-attention sub-network is added to help the current node obtain the current content which requires to be focused.

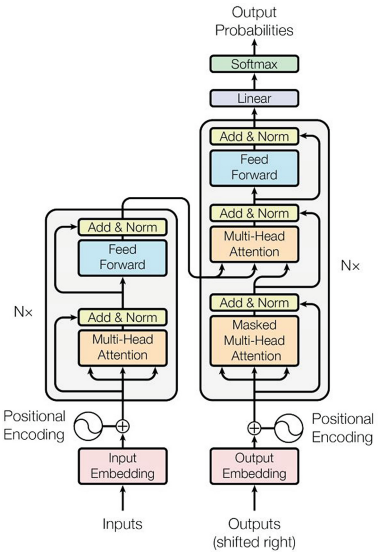

Fig. 2. The architecture of Transformer, where the encoder is made up of six layers with each contains two sub-networks, namely multi-head self-attention mechanism (MSHA) and a fully connected network layer.

Backbone Network with Attention. BERT [5] mainly consists of a stack of multiple Transformer encoders and uses a bidirectional structure to handle classification or regression tasks. With its pre-trained models, we can fine-tune in cross-modal retrieval tasks by learning a more comprehensive textual feature representation. In this work, we also compared bi-LSTM. Extensive experiments show that BERT as the textual backbone network has a great improvement.

Although CNN can effectively capture local information, long-distance dependencies need to be established in visual tasks such as object detection, instance segmentation and key point detection. Vision Transformer (ViT) [6] shows that the dependence of computer vision tasks on CNN is unnecessary, and the direct application of image block sequence converters can perform image classification tasks well. The specific method is to split the image into small blocks

and provide the linear embedding sequence of these small blocks as the input. The specific structure is shown in Fig. 3.

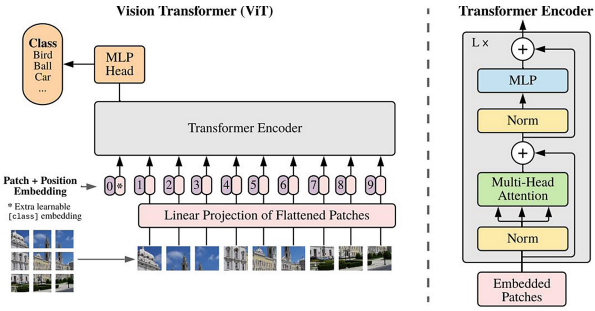

Fig. 3. ViT framework. The image with $(\mathrm{H} \times \mathrm{W} \times \mathrm{C})$ is flattened into a two-dimensional block, whose size is $(N \times (P^2C))$. Among them, (P, P) is the size of the block, $N = HW/P^2$, where N is block number. Flatten tiles are the input of the encoder and the structure of the decoder is similar to Transformer.

Although ViT first applied Transformer to computer vision, compared with image classification, the size of image is usually larger, hence the application is limited. Besides, dividing images into tiles and expanding them into a line will damage the continuity of information in the horizontal image space. Specifically, the two tiles arranged vertically will be separated by the others after being expanded.

In order to solve these problems, [16,22,29] add a self-attention layer to CNN layers. While fully expressing the local representation, it also pays attention to long sequential information, which makes some substantial progress. [20] designed the following ideas: 1) Preserve the CNN layers to effectively learn coarse feature maps from the image; 2) Design global-grained attention layer to integrate the information captured by CNN. In this way, the advantages of Transformer and CNN are combined and spatial downsampling is performed through convolution so that the attention layer can work more effectively at images with a smaller resolution. In this work, we refer to [20], replace the last three 3×3 convolutional layers of ResNet with MHSA, as is shown in Fig. 3 and 4.

Bottleneck Transformer as the image backbone network learns better image feature representation. In the experiment, we also compared it with the traditional ResNet, which proves Bottleneck Transformer indeed achieved a certain improvement.

3.3 Visual and Textual Semantic Alignment

The image backbone is denoted as $F(x) = BoTNet(x)$, where $x \in (64, 3, 384, 128)$ which represent the batch size, channels and the dimension of the

image respectively. Next, we introduce global-grained branch and fine-grained branch, as shown in Fig. 2. $Global = F_{global}(F(x))$, each local feature of fine-grained branch can be expressed as $Fine = F_{fine}(F(x))$.

A segmentation layer is set refer to [26] with fine-grained branches and local attribute representation obtained through the Human parsing network. Each fine-grained branch is supervised to create a segmentation map of a concrete attribute searching list. It can be treated as a knowledge regulator to present the attributes of each fine-grained branch.

The next stage is to learn an associative embedding between two modalities. We introduce contrastive learning and treat the input as a triplet, namely $\langle visual^i, textual^*, textual \rangle$ and $\langle textual^i, visual^*, visual \rangle$, where i means the anchor to be identified, * represents the feature representation corresponding to the i^{th} person, namely a positive sample. The rest represents a randomly selected person irrelevant to the i^{th} person, namely a negative sample. For matching evaluation method, We use cosine similarity to evaluate, which can be specifically expressed as:

$$S_{cosine} = \frac{v_t \cdot t}{\|v\| \cdot \|t\|} \tag{4}$$

For a positive sample $\langle v^i, t^+ \rangle$, larger the similarity, the better. For negative samples, the similarity should be as small as possible. However, simply suppressing the similarity will have constraints on negative samples, so we choose to set a boundary. The goal is to optimize the deviation between S^+_{cosine} and S^-_{cosine} to be greater than the boundary value, which is also called Relative similarity criterion. It can be denoted as the follows:

$$S^+_{cosine} \to 1 \tag{5}$$

$$S^+_{cosine} - S^-_{cosine} > m \tag{6}$$

We select logistic loss as the loss function, hence the equation above can be simplified as:

$$S^+_{cosine} - \alpha > 0 \tag{7}$$

$$S^+_{cosine} - \beta < 0 \tag{8}$$

α is the lower boundary of a positive sample, β is the upper boundary of a negative sample. According to logistic function, the ultimate objective function is:

$$L_{AlignLoss} = \frac{1}{N} \sum_{i=1}^{N} log[1 + e^{-\tau_r(S_i^+ - \alpha)}] + log[1 + e^{-\tau_j(S_i^- - \beta)}] \tag{9}$$

where τ_r and τ_j represents two hyper parameters to control size of the gradient.

One of the prerequisites for correctly retrieving image and text pairs is a large number of positive and negative samples, which can provide effective supervised learning. However, the current mainstream contrastive learning methods select samples belonging to the same category, while only arbitrarily choose negative samples, which has little effect on the matching of global features but suffers from local features. Therefore, to process the information expressed in fine-grained

branch, we need to expand the searching space of positive samples to construct a more comprehensive sample pair.

When two images with the same attributes belong to different identities, they can be called "surrogate positive sample". For such samples, how to deal with fine-grained semantic difference is the key point. Subject to the rearrangement technology in the Re-ID, k-reciprocal unsupervised sampling [26] learns to gain attribute-level labels effectively. Specifically, we can extract a batch of visual and textual feature representations from backbone and use k-reciprocal unsupervised sampling to mine the corresponding "surrogate positive samples" of each attribute.

For the instance level, we add cross-entropy loss separately to help learn their discriminative features, that is, the Instance loss. Pixel-level cross-entropy loss is added for the local attribute features obtained through the Human parsing network, which is the Segmentation loss. What's more, in visual-textual alignment, we designed matching loss function in global granularity and fine granularity, namely Global-grained loss and Fine-grained loss. The overall objective function can be specifically denoted as:

$$L_{overall} = L_{instance} + L_{segmentation} + L_{global} + L_{fine} \tag{10}$$

4 Experiments and Results

4.1 Evaluation Dataset

We evaluated the proposed methods on CUHK-PEDES, which is currently the only cross-modal pedestrian retrieval dataset based on natural language description. It consists of 40,206 pictures and 13,003 identities in total. The training set is composed of 34054 images, including 11,003 identities; the validation set is composed of 3078 images, including 1,000 identities; and the test set is composed of 3074 images, including 1,000 identities and at least two descriptions for each.

4.2 Experiment Setup

All the experiments are based on PyTorch. The model used in the experiments was implemented on NVIDIA Tesla V100 and NVIDIA GeForce RTX 3080 GPU. We use Recall@K to evaluate the performance, which means that in the top-K results, at least the percentage of the text description corresponding to the image of a matching person is retrieved.

For images, we use the model shown in Fig. 3 as the backbone, and ResNet-X for comparison. The input image size is 384×128. Adam optimizer and weight decay is set to 4×10^{-5}. Each mini-batch contains 64 image-text pairs. The learning rate is initialized to 2×10^{-4} in the first 40 iterations, and then decay by 0.1 times in the remaining 30 iterations. The hyper parameters of formula 9 are set to $\alpha = 0.6, \beta = 0.4, \tau_r = 10, \tau_j = 40$, respectively.

For attribute level, we use the Human parsing network to divide the visual-textual attributes into five categories: head, upper body, lower body, shoes, and

bags. These five categories can describe the appearance of a pedestrian in detail. In Fig. 4, we visualize the body segmentation result, which proves effective performance of our method where different color blocks represent different feature attributes.

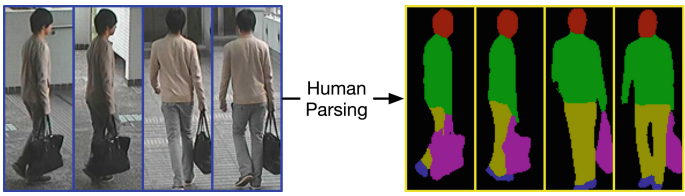

Fig. 4. Segmentation of 5 human body parts

4.3 Comparisons with *State-of-the-Art*

We summarized the global and fine-grained representations, and compared them with the current *state-of-the-art* methods under CUHK-PEDES by using Recall@ K (K = 1, 5, 10). Dual Path [9], CMPM+CMPC [28] designed better joint embeddings to learn the relationship between images and text, while GNA-RNN [?], CMCE [11] and PWM-ATH [4] use attention mechanism to express the relationship between images and text. Without exception, these methods use global representation learning. Compared with ViTAA, we have obtained a more adequate global and fine-grained feature representation after optimizing the backbone network of both images and text, which has improved in all indicators. The specific test results are shown in Table 1.

Table 1. Comparison with *state-of-the-art* methods on CUHK-PEDES

Methods	Feature selection	R@1	R@5	R@10
GNA-RNN	Global	19.05	–	53.64
CMCE	Global	25.94	–	60.48
CMPM+CMPC	Global	49.37	–	79.27
Dual Path	Global	44.40	66.26	75.07
ViTAA	Global+Fine-grained	55.97	75.84	83.52
Proposed method	Global+Fine-grained	**57.98**	**77.37**	**84.84**

The proposed method has improved in all indicators, R@1 is 2% higher than ViTAA, while R@5 is 1.53% higher and R@10 is 1.32% higher.

4.4 Ablation Studies

After analyzing the overall performance of our proposed model, we test each part of the model and use the test results to verify the role of each part. The specific test details are as follows. To test the performance of each part, We still used CUHK-PEDES, and the model is divided into four parts: image backbone network, textual backbone network, global-grained module and fine-grained module.

Image Backbone Network. For the image backbone, we use ResNet-X and our proposed method to test the feature extraction capabilities, while keeping the text backbone network all set to BERT, and the global-grained branch and the fine-grained branch are all retained, The test details are shown in Table 2.

Table 2. Comparison of different image backbones

Backbone selection		R@1	R@5	R@10
Visual	Textual			
ResNet-18	BERT	46.60	69.24	78.36
ResNet-34	BERT	55.96	74.18	84.24
ResNet-50	BERT	55.99	75.84	83.51
Proposed method	BERT	**57.98**	**77.37**	**84.84**

From the table above, the proposed image backbone has achieved the highest retrieval performance under while other modules remain unchanged. The fusion of CNN and MHSA can better extract visual feature representation.

Textual Backbone Network. For the textual backbone, we use LSTM, Bi-LSTM and our proposed method to test the textual feature extraction capability, while keeping the image backbone network all set to BotNet-50, and the global and fine-grained branches are all retained. Test details are shown in Table 3.

Table 3. Ablation studies of different textual backbone network

Backbone selection		R@1	R@5	R@10
Textual	Visual			
LSTM	BotNet-50	55.46	75.14	78.35
bi-LSTM	BotNet-50	55.98	75.85	83.55
Proposed method	BotNet-50	**57.98**	**77.37**	**84.84**

From the results, we can conclude that our textual backbone has achieved better retrieval results than the others, which proves BERT is better than traditional models in textual feature extraction. However, BERT is relatively larger in framework and requires higher computing power and hardware conditions. Compared with LSTM, BERT does better in processing long-sequence information.

Visual-Textual Alignment. The visual-textual alignment can be divided into the following parts: Instance loss, Segmentation loss, Global-grained loss and Fine-grained loss. In order to compare the contribution of each part of the module, we keep one part and remove all the other parts to compare the effect of the individual modules. Test results are shown in Table 4.

Table 4. Ablation studies of visual-textual alignment module

Model selection				R@1	R@5	R@10
Instance	Segmentation	Global	Fine			
*				29.52	51.82	61.57
*	*			30.39	52.71	63.11
*	*		*	32.16	53.91	63.67
*	*	*		54.81	74.02	82.45
*	*	*	*	**57.98**	**77.37**	**84.84**

The results in Table 4 certify that: 1) The purpose of fine-grained loss is to improve the marginal optimization of local feature, which only has a feeble help to the matching process. Similarly, when the segmentation loss is detached, the local attribute feature cannot be fully aligned, leading to a decrease in global-grained module. It can be seen that the segmentation loss and the global-grained loss are indispensable. 2) When both the fine-grained loss and segmentation loss are selected, the performance is greatly improved, indicating that segmentation loss and global-grained loss complement each other well. 3) The last two rows compare the impact of fine-grained loss. It can be seen that the global-grained loss contributes the most, which greatly improves the retrieval effect. After integrating all modules, the performance has been slightly improved, which proves the contribution of fine-grained loss again.

5 Conclusion

In this paper, we design a cross-modal retrieval framework for person search by natural language, adding attention mechanism to both image and textual backbones to more fully discover the features from these modalities. At the same time, on the basis of retaining the CNN layer in the image backbone network, we

apply Transformer so as to effectively learn features at attribute level and low-resolution maps from the images, while also using the global self-attention layer to roll the information captured from CNN layers. In this way, the advantages of Transformer and CNN are combined, and spatial downsampling is performed in order that the attention layer can work more effectively at a smaller resolution. Besides, global and fine grained branches contribute to the performance a lot, which proves the idea that both feature from global-grained level and attribute level deserves to be discovered. Experiments on CHUK-PEDES demonstrate that our method outperforms *state-of-the-arts* results. In the future, we will focus on the generalization of this work with tests on other datasets to further elaborate the framework.

References

1. Andrew, G., Arora, R., Bilmes, J., Livescu, K.: Deep canonical correlation analysis. In: International Conference on Machine Learning, pp. 1247–1255. PMLR (2013)
2. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)
3. Cai, G., et al.: Ask&confirm: active detail enriching for cross-modal retrieval with partial query. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 1835–1844, October 2021
4. 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)
5. 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)
6. Dosovitskiy, A., et al.: An image is worth 16×16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
7. 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)
8. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
9. Huang, Y., Wang, W., Wang, L.: Instance-aware image and sentence matching with selective multimodal LSTM. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2310–2318 (2017)
10. Jeon, J., Lavrenko, V., Manmatha, R.: Automatic image annotation and retrieval using cross-media relevance models. In: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 119–126 (2003)
11. 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)
12. Liu, Y., Guo, Y., Bakker, E.M., Lew, M.S.: Learning a recurrent residual fusion network for multimodal matching. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4107–4116 (2017)
13. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**(2), 91–110 (2004)

14. Ma, L., Lu, Z., Shang, L., Li, H.: Multimodal convolutional neural networks for matching image and sentence. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2623–2631 (2015)
15. Nam, H., Ha, J.W., Kim, J.: Dual attention networks for multimodal reasoning and matching. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 299–307 (2017)
16. Ramachandran, P., Parmar, N., Vaswani, A., Bello, I., Levskaya, A., Shlens, J.: Stand-alone self-attention in vision models. arXiv preprint arXiv:1906.05909 (2019)
17. Reed, S., Akata, Z., Lee, H., Schiele, B.: Learning deep representations of fine-grained visual descriptions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 49–58 (2016)
18. Schmidhuber, J.: Deep learning in neural networks: an overview. Neural Netw. **61**, 85–117 (2015)
19. Schroff, F., Kalenichenko, D., Philbin, J.: FaceNet: a unified embedding for face recognition and clustering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 815–823 (2015)
20. 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)
21. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 5998–6008 (2017)
22. Wang, H., Zhu, Y., Green, B., Adam, H., Yuille, A., Chen, L.-C.: Axial-DeepLab: stand-alone axial-attention for panoptic segmentation. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12349, pp. 108–126. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58548-8_7
23. Wang, K., Yin, Q., Wang, W., Wu, S., Wang, L.: A comprehensive survey on cross-modal retrieval. arXiv preprint arXiv:1607.06215 (2016)
24. Wang, L., Li, Y., Huang, J., Lazebnik, S.: Learning two-branch neural networks for image-text matching tasks. IEEE Trans. Pattern Anal. Mach. Intell. **41**(2), 394–407 (2018)
25. Wang, L., Li, Y., Lazebnik, S.: Learning deep structure-preserving image-text embeddings. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5005–5013 (2016)
26. 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
27. Yan, F., Mikolajczyk, K.: Deep correlation for matching images and text. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3441–3450 (2015)
28. Zhang, Y., Lu, H.: Deep cross-modal projection learning for image-text matching. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 686–701 (2018)
29. Zhao, H., Jia, J., Koltun, V.: Exploring self-attention for image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10076–10085 (2020)
30. Zheng, Z., Zheng, L., Garrett, M., Yang, Y., Xu, M., Shen, Y.D.: Dual-path convolutional image-text embeddings with instance loss. ACM Trans. Multimed. Comput. Commun. Appl. (TOMM) **16**(2), 1–23 (2020)