LEARNING SEMANTIC-ALIGNED FEATURE REPRESENTATION FOR TEXT-BASED PERSON SEARCH

Shiping Li[1], *Min Cao*[1*], *Min Zhang*[2]

[1]Soochow University, [2]Harbin Institute of Technology, Shenzhen
spli@stu.suda.edu.cn, {mcao,minzhang}@suda.edu.cn

ABSTRACT

Text-based person search aims to retrieve images of a certain pedestrian by a textual description. The key challenge of this task is to eliminate the inter-modality gap and achieve the feature alignment across modalities. In this paper, we propose a semantic-aligned embedding method for text-based person search, in which the feature alignment across modalities is achieved by automatically learning the semantic-aligned visual features and textual features. First, we introduce two Transformer-based backbones to encode robust feature representations of the images and texts. Second, we design a semantic-aligned feature aggregation network to adaptively select and aggregate features with the same semantics into part-aware features, which is achieved by a multi-head attention module constrained by a cross-modality part alignment loss and a diversity loss. Experimental results on the CUHK-PEDES and Flickr30K datasets show that our method achieves state-of-the-art performances.

Index Terms— Text-based person search, semantic alignment, multi-head attention, Transformer

1. INTRODUCTION

Text-based person search [1] aims to search for the corresponding person images from a large-scale image database given a textual description. Its challenge lies in two aspects: feature extraction from both visual and textual modalities and cross-modal alignment. First, it is challenging to extract robust feature representations from both images and texts due to background clutter, pose/viewpoint variances in the images, and the complexity of natural language. Then, it is difficult to overcome the cross-modal gap for alignment.

Various text-based person search methods have been proposed in recent years. We generally categorize them into global-matching methods and local-matching methods. Global-matching methods [2, 3, 1, 4, 5] extract the global representation of samples from the two modalities separately and design proper objective functions to explore a shared latent embedding space, in which the matching scores for image-text pairs can be computed directly. However, the global-matching methods cannot effectively explore the distinctive local details of samples, which are beneficial to improving the performance.

To further mine discriminative and comprehensive information, local-matching methods are proposed. The existing local-matching methods [6, 7, 8, 9, 10, 11, 12] generally consist of two procedures: local feature extraction and cross-modal alignment. (1) For local feature extraction, the local units in images and texts are firstly obtained by pre-set rules, based on which the local features from the local units are explicitly extracted. Specifically, some methods [11, 7] obtain the local units by simply dividing the image or its feature map into stripes or patches, and dividing the text into words, and then compute the local features by directly extracting feature representations from these units. This way introduces background noise into the features, which heavily influences the following alignment phase and causes sub-optimal retrieval results. For this reason, some works leverage additional models, such as human parsing [6], pose estimation [10], and attribute recognition [9], to locate the semantic parts in image and text as local units, yet with a heavy computing burden and non end-to-end architecture. (2) For cross-modal alignment, most local-matching methods [10, 11, 13, 12] adopt the cross-modality attention mechanism to explore the alignment between visual local units and textual ones.

To sum up, global-matching methods roughly learn a global representation by jointly embedding the images and texts into a shared space. Local-matching methods align the local units through a cross-modality attention mechanism and the local units are obtained by simple dividing operations or extra models. Compared with global-matching methods, local-matching methods can significantly boost performance owing to the information exploration at a fine-grained level and the information interaction between modalities. However, the information interaction in local-matching methods inevitably brings about efficiency damage at inference and is hardly practical in real-life applications. Thus, it is necessary to develop a simple but effective method for text-based person search, which this paper focuses on.

Motivated by the above observation, we propose a novel semantic-aligned embedding method for text-based person

*Corresponding author

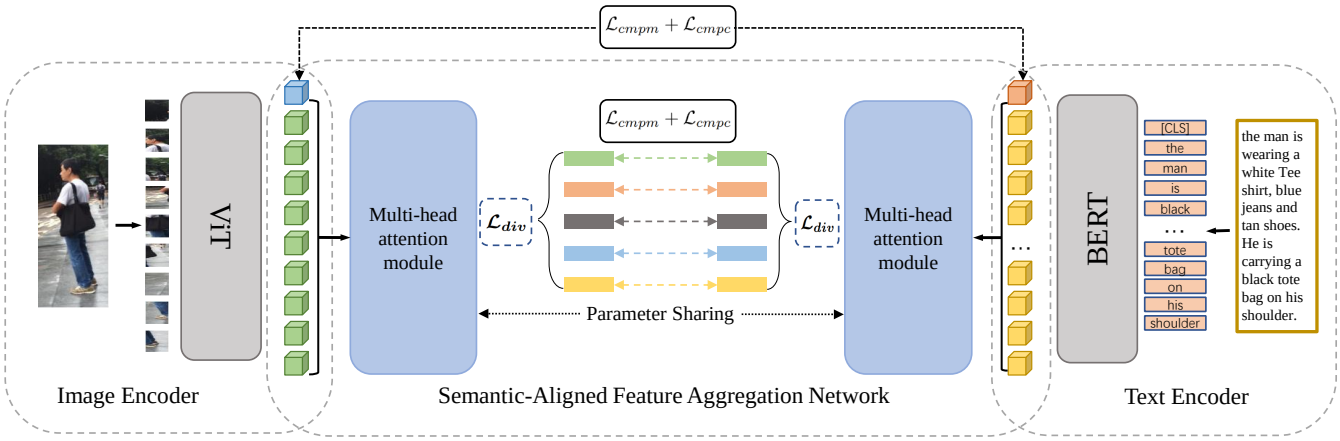

Fig. 1: The overall architecture of the proposed method.

search, in which the semantic-aligned part-aware features across modalities can be automatically obtained without extra model intervention and cross-modality attention mechanism. Our main contributions are three folds: (1) We propose a semantic-aligned feature aggregation network, which adaptively aggregates unit features with the same semantic into diverse part-aware features. (2) To the best of our knowledge, we are the first to introduce transformer-based backbones ViT [14], and BERT [15] into both visual modality and textual modality to extract robust feature representations in the text-based person search. (3) The experimental results on CUHK-PEDES [1], and Flickr30K [16] achieve state-of-the-art performances, which verifies the effectiveness and generalization of the proposed method.

2. METHOD

Fig.1 shows an overview of the proposed method that includes the modality-specific feature encoders (*i.e.*, an image encoder and a text encoder) and a semantic-aligned feature aggregation network.

2.1. Modality-specific Feature Encoder

Image Encoder. We adopt a ViT [14] pretrained on ImageNet [17] as the image feature encoder. Given an image, we first split it into a sequence of N fixed-sized patches grid-likely, and then map the patch sequence to d dimensional embeddings by a trainable linear projection. An extra learnable [IMG] embedding token is prepended to the sequence of patch embeddings to learn global representation. Then, we feed the patch embeddings into Transformer encoder. The output is denoted as $\mathbf{E} = \{\mathbf{e}_g, \mathbf{e}_1, \cdots, \mathbf{e}_N\} \in \mathbb{R}^{(N+1)\times d}$, where $\mathbf{e}_g$ is the global feature of the input image, $\{\mathbf{e_1}, \cdots, \mathbf{e}_N\}$ are patch features.

Text Encoder. Given a textual description with M words, we feed the sequence of words to a pretrained BERT [15] to extract the textual features $\mathbf{T} = \{\mathbf{t}_g, \mathbf{t}_1, \cdots, \mathbf{t}_M\} \in \mathbb{R}^{(M+1)\times d}$, where $\mathbf{t}_g$ is the global feature of the input text from an extra [CLS] token, $\{\mathbf{t}_1, \cdots, \mathbf{t}_M\}$ are word features.

2.2. Semantic-Aligned Feature Aggregation Network

The similarity of the image-text pair can be computed by aligning the patches in the image and the words in the text, *i.e.*, direct comparison between $\mathbf{E}$ and $\mathbf{T}$. However, such similarity is unreliable for text-based person search due to background noise in patches and words. For this reason, we compute the similarity by aligning the image regions and the corresponding phrases in the text at the semantic level. At that point, we propose a semantic-aligned feature aggregation network, in which the region-phrase alignment is achieved by using a multi-head attention [18] module constrained by a cross-modality part alignment loss and a diversity loss.

Multi-head attention module. For the visual modality, the multi-head attention module takes $\mathbf{E}$ as input and output K embeddings, each of which is a weighted sum of patch features. Concretely, inputting the visual features $\mathbf{E}$, we firstly calculate three vectors in i-th head ($i = 1, \cdots, K$): query $\mathbf{Q}_i$, key $\mathbf{K}_i$, and value $\mathbf{V}_i$ through the linear projections,

$$\mathbf{Q}_i = \mathbf{E}\mathbf{W}_i^Q, \mathbf{K}_i = \mathbf{E}\mathbf{W}_i^K, \mathrm{V}_i = \mathbf{E}\mathbf{W}_i^V, \quad (1)$$

where the trainable parameter matrices $\mathbf{W}_i^Q, \mathbf{W}_i^K, \mathbf{W}_i^V \in \mathbb{R}^{d\times d}$. Then we compute the i-th attention weight matrix $\mathbf{A}_i \in \mathbb{R}^{(N+1)\times(N+1)}$ of the input image as

$$\mathbf{A}_i = \mathrm{softmax}(\frac{\mathbf{Q}_i\mathbf{K}_i^T}{\sqrt{d}}), \quad (2)$$

Based on this, we obtain K embeddings through

$$\mathbf{E}_i = \mathbf{A}_i\mathbf{V}_i. \quad (3)$$

In i-th head, $\tilde{\mathbf{e}}_i = \mathbf{E}_i(0,:) \in \mathbb{R}^d$ represents the global representation encoded by this head, which is a weighted sum of patch features. Therefore, we obtain the visual embedding set $\tilde{\mathbf{E}} = \{\tilde{\mathbf{e}}_1, \cdots, \tilde{\mathbf{e}}_K\} \in \mathbb{R}^{K\times d}$ containing K feature representations of the input image encoded by different heads.

In the textual modality, we input the $\mathbf{T}$ to the multi-head attention module that shares the same parameters with that in

the visual modality, and output K textual embeddings $\tilde{\mathbf{T}} = \{\tilde{\mathbf{t}}_1, \cdots, \tilde{\mathbf{t}}_K\} \in \mathbb{R}^{K \times d}$.

Cross-modality Part Alignment. To achieve region-phrase semantic alignment, we constrain a one-to-one alignment relationship between visual embeddings $\tilde{\mathbf{E}}$ and textual embeddings $\tilde{\mathbf{T}}$ by introducing a cross-modal part alignment loss, including a Cross-Modal Projection Matching (CMPM) loss and a Cross-Modal Projection Classification (CMPC) loss [5].

For a batch containing n image-text pairs, through the multi-head attention module, we obtain K visual embedding matrices denoted by $\left\{\tilde{\boldsymbol{\mathcal{E}}}_1, \cdots, \tilde{\boldsymbol{\mathcal{E}}}_K\right\}$, where $\tilde{\boldsymbol{\mathcal{E}}}_k \in \mathbb{R}^{n \times d}$, and K textual embedding matrices denoted by $\left\{\tilde{\boldsymbol{\mathcal{T}}}_1, \cdots, \tilde{\boldsymbol{\mathcal{T}}}_K\right\}$. The cross-modality part alignment loss is defined as

$$\mathcal{L}_{part} = \frac{1}{K}\sum_{k}^{K}(\mathcal{L}_{cmpm}(\tilde{\boldsymbol{\mathcal{E}}}_k, \tilde{\boldsymbol{\mathcal{T}}}_k) + \mathcal{L}_{cmpc}(\tilde{\boldsymbol{\mathcal{E}}}_k, \tilde{\boldsymbol{\mathcal{T}}}_k)), \quad (4)$$

where the $\mathcal{L}_{cmpm}$ utilizes the KL divergence to minimize the similarities between texts and images with the different identities while maximizing the similarities between texts and images with the same identity, and the $\mathcal{L}_{cmpc}$ encourages the feature representations of samples with the same identity to be similar and discriminate from feature representations of other samples with the different identity. As a result, the visual feature $\tilde{\mathbf{e}}_k$ and the textual feature $\tilde{\mathbf{t}}_k$ with the same identity in k-th head contain the same semantic information.

Diversity Regularization. Considering that different head attention blocks could capture the redundant and overlapped semantic information to each other in the multi-head attention module, we take a further step to introduce a diversity loss $\mathcal{L}_{div}$ that penalizes the redundancy,

$$\mathcal{L}_{div} = \frac{1}{K(K-1)}\sum_{i=1}^{K}\sum_{j=1, i\neq j}^{K}\left(\frac{\tilde{\mathbf{e}}_i \cdot \tilde{\mathbf{e}}_j}{\|\tilde{\mathbf{e}}_i\|_2\|\tilde{\mathbf{e}}_j\|_2} + \frac{\tilde{\mathbf{t}}_i \cdot \tilde{\mathbf{t}}_j}{\|\tilde{\mathbf{t}}_i\|_2\|\tilde{\mathbf{t}}_j\|_2}\right) \quad (5)$$

Therefore, the embeddings in $\tilde{\mathbf{E}}$ and $\tilde{\mathbf{T}}$ can represent different part of the input image and the input text, respectively. Through the above cross-modality part alignment loss and diversity loss, we obtain diverse semantic-aligned part-aware features from the image and text, *i.e.*, $\tilde{\mathbf{E}}$ and $\tilde{\mathbf{T}}$.

2.3. Training and inference

The entire network is trained in an end-to-end manner. As a supplementary to the cross-modal part alignment, we add a cross-modal global alignment. We denote the global features of images and texts in a batch as $\boldsymbol{\mathcal{E}}_g \in \mathbb{R}^{n \times d}$ and $\boldsymbol{\mathcal{T}}_g \in \mathbb{R}^{n \times d}$, respectively. The overall loss is computed by

$$\mathcal{L} = \mathcal{L}_{global} + \mathcal{L}_{part} + \lambda\mathcal{L}_{div}, \quad (6)$$

where $\mathcal{L}_{global} = \mathcal{L}_{cmpm}(\boldsymbol{\mathcal{E}}_g, \boldsymbol{\mathcal{T}}_g) + \mathcal{L}_{cmpc}(\boldsymbol{\mathcal{E}}_g, \boldsymbol{\mathcal{T}}_g)$ and λ is a parameter to control the importance of the diversity loss.

During the inference, the similarity score between one image-text pair is measured as the sum of the cosine distance of the global features and part-aware features between them,

$$Sim(I, T) = cosine(\mathbf{e}_g, \mathbf{t}_g) + \frac{1}{K}\sum_{k}^{K} cosine(\tilde{\mathbf{e}}_k, \tilde{\mathbf{t}}_k). \quad (7)$$

3. EXPERIMENTS

3.1. Datasets and implementation details

We conduct experiments on the only text-based person search benchmark CUHK-PEDES [1], which contains 40,206 images of 13,003 identities, with 2 captions per image, 11,003 for training, 1,000 for evaluation, and 1,000 for test. To verify the generalization of the proposed method, we also conduct experiments on a cross-modal retrieval dataset Flickr30K [16], in which there are 31,784 images with five captions for each in total. We take the same split for training, validation and testing set as [19]. Flickr30K contains a variety of objects, rather than only pedestrians as CUHK-PEDES.

In the experiments, we resize the image to 224×224 and pad the text to length 100. There are 50 epochs in training, and the batch size is set to 64. The dimension of visual and textual features is set to $d = 768$. The parameter K for the multi-head attention module is set to 10 and the parameter λ in Eq. 6 is set to 0.2. Following the standard setting [6], we adopt Rank-1/5/10 as the evaluation criteria.

3.2. Comparisons with State-of-the-art Methods

Results on CUHK-PEDES Dataset. The comparison results in Table 1 show that the proposed method performs the best results at Rank-1/5/10 accuracies with all comparisons. Particularly, the proposed method outperforms the second-best method SSAN [20] by a large margin, such as a 2.76% gain at Rank-1. It is worth noting that the compared methods GLA [12], PMA[10] and MIA [11] construct the elaborate cross-modality attention mechanism for achieving cross-modal alignment with low efficiency. By contrast, the proposed method reaches overbearing advantages on performance by performing an automatic cross-modal alignment with a simple and lightweight architecture.

Results on Flickr30K Dataset: We compare the proposed method against several state-of-the-art methods in Table 2. It can be shown that the proposed method surpasses all other methods at all Rank accuracies, showing the effectiveness and generalization of the proposed method.

3.3. Ablation Studies

Analysis on modules of the proposed method: There are two modules in the proposed method: the modality-specific feature encoder (MSFE) and the semantic-aligned feature aggregation network (SAFA). The global matching from the first module and the local matching from the second one is incorporated into the matching computation for text-based person search. We analyze the impacts of the two modules on

Table 1: Performance comparison with state-of-the-arts on CUHK-PEDES dataset.

Method	Rank-1	Rank-5	Rank-10
GNA-RNN [1]	19.05	-	53.64
GLA [12]	43.58	66.93	76.26
Dual-path [3]	44.40	66.26	75.07
CMPM+CMPC [5]	49.27	-	79.27
MCCL [2]	50.58	-	79.06
MIA [11]	53.10	75.00	82.90
A-GANet [13]	53.14	74.03	81.95
PMA [10]	53.81	73.54	81.23
TIMAM [21]	54.51	77.56	84.78
ViTAA [6]	55.97	75.84	83.52
MGEL [8]	60.27	80.01	86.74
SSAN [20]	61.37	80.15	86.73
Ours	**64.13**	**82.62**	**88.40**

Table 2: Performance comparison with state-of-the-arts on Flickr30K dataset.

Method	Text-to-Image		
	Rank-1	Rank-5	Rank-10
Dual-path [3]	39.10	69.20	80.90
CMPM+CMPC [5]	37.30	65.70	75.50
A-GANet [13]	39.52	69.91	80.91
TIMAM [21]	42.60	71.60	81.90
Ours	**50.74**	**77.92**	**85.46**

performance. The results are shown in Table 3. Compared to Ours$_g$, Ours achieves performance enhancement, (*i.e.*, 4.34%, 3.02% and 2.39% gains at Rank-1/5/10), which demonstrates the effectiveness of the proposed local-matching from the SAFA module. In addition, Ours obtains the minor increases at Rank-1/5 compared with ours$_p$, indicating that the global-matching from the MSFE module is an effective auxiliary for the cross-modal matching.

Parameters analysis: There are two parameters K and λ in the proposed method. Fig.2 shows the impact of two parameters at Rank-1. (1) For parameter K, the best result is achieved with $K = 10$. A smaller value to K represents insufficient exploration on the fine-grained region-phrase alignment and results in performance degradation. Besides, a larger value to K may introduce noise information for alignment and lead to the decrease of performance. (2) For parameter λ, the proposed method with $\lambda = 0.2$ performs the best result. We can study that when setting $\lambda > 0$, the proposed method always achieves performance enhancement, showing the effectiveness of the diversity loss for performance.

Table 3: Ablation study of the proposed method on CUHK-PEDES dataset.

Method	global	part	Rank-1	Rank-5	Rank-10
Ours$_g$	✓		59.80	79.60	86.01
Ours$_p$		✓	63.90	82.50	**88.71**
Ours	✓	✓	**64.13**	**82.62**	88.40

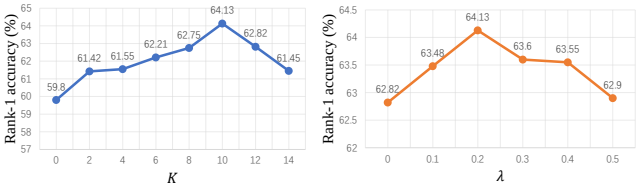

Fig. 2: Influence of the parameters in the proposed method on CUHK-PEDES.

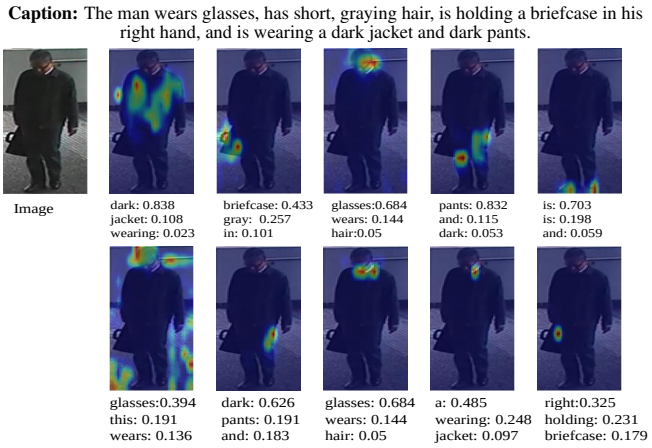

Fig. 3: Visualization of the semantic alignment across modalities by the proposed method.

3.4. Visualization Analysis

We visualize the semantic alignment across modalities from the multi-head attention module in the proposed method. Specifically, we show the attention map of each head in the image and the words with the first three attention values among all the words in the text in Fig.3. We can observe that each head attends to a different part of the image and the text, and the part's semantic information from them is interrelated with each other. It shows that the proposed method achieves effective semantic alignment across modalities.

4. CONCLUSIONS

In this paper, we propose a simple but effective method for text-based person search. In contrast to the existing local-matching methods, the proposed method is an end-to-end trainable architecture without additional models and complex cross-modal information interaction strategies. We design a semantic-aligned feature aggregation network for learning the part-aware features that are semantic-aligned across modalities adaptively. The experimental results on CUHK-PEDES and Flickr30K verify the superiority of the proposed method.

Acknowledgement. This work is supported by the National Science Foundation of China under Grant NSFC 62002252, and is also partially supported by Collaborative Innovation Center of Novel Software Technology and Industrialization.

5. REFERENCES

[1] Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang, "Person search with natural language description," in *CVPR*, 2017, pp. 1970–1979.

[2] Yuyu Wang, Chunjuan Bo, Dong Wang, Shuang Wang, Yunwei Qi, and Huchuan Lu, "Language person search with mutually connected classification loss," in *ICASSP*. IEEE, 2019, pp. 2057–2061.

[3] 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.

[4] Shuang Li, Tong Xiao, Hongsheng Li, Wei Yang, and Xiaogang Wang, "Identity-aware textual-visual matching with latent co-attention," in *ICCV*, 2017, pp. 1890–1899.

[5] Ying Zhang and Huchuan Lu, "Deep cross-modal projection learning for image-text matching," in *ECCV*, 2018, pp. 686–701.

[6] Zhe Wang, Zhiyuan Fang, Jun Wang, and Yezhou Yang, "Vitaa: Visual-textual attributes alignment in person search by natural language," in *ECCV*. Springer, 2020, pp. 402–420.

[7] Kecheng Zheng, Wu Liu, Jiawei Liu, Zheng-Jun Zha, and Tao Mei, "Hierarchical gumbel attention network for text-based person search," in *ACM MM*, 2020, pp. 3441–3449.

[8] Chengji Wang, Zhiming Luo, Yaojin Lin, and Shaozi Li, "Text-based person search via multi-granularity embedding learning," in *IJCAI*, 2021.

[9] Surbhi Aggarwal, Venkatesh Babu RADHAKRISHNAN, and Anirban Chakraborty, "Text-based person search via attribute-aided matching," in *WACV*, 2020, pp. 2617–2625.

[10] Ya Jing, Chenyang Si, Junbo Wang, Wei Wang, Liang Wang, and Tieniu Tan, "Pose-guided multi-granularity attention network for text-based person search," in *AAAI*, 2020, vol. 34, pp. 11189–11196.

[11] 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.

[12] Dapeng Chen, Hongsheng Li, Xihui Liu, Yantao Shen, Jing Shao, Zejian Yuan, and Xiaogang Wang, "Improving deep visual representation for person re-identification by global and local image-language association," in *ECCV*, 2018, pp. 54–70.

[13] Jiawei Liu, Zheng-Jun Zha, Richang Hong, Meng Wang, and Yongdong Zhang, "Deep adversarial graph attention convolution network for text-based person search," in *ACM MM*, New York, NY, USA, 2019, p. 665–673, Association for Computing Machinery.

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2021.

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[16] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 67–78, 2014.

[17] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[19] Andrej Karpathy and Li Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3128–3137.

[20] 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.

[21] Zheng-Jun Zha, Jiawei Liu, Di Chen, and Feng Wu, "Adversarial attribute-text embedding for person search with natural language query," *IEEE Transactions on Multimedia*, vol. 22, no. 7, pp. 1836–1846, 2020.