See Finer, See More: Implicit Modality Alignment for Text-based Person Retrieval

Xiujun Shu[1⋆], Wei Wen[1⋆], Haoqian Wu[1], Keyu Chen[1], Yiran Song[1,2],
Ruizhi Qiao[1], Bo Ren[1], Xiao Wang[3†]

[1]Tencent Toutu Lab
{xiujunshu, jawnrwen, linuswu, yolochen, ruizhiqiao, timren}@tencent.com
[2]Shanghai Jiao Tong University [3]Anhui University
songyiran@sjtu.edu.cn, wangxiaocvpr@foxmail.com

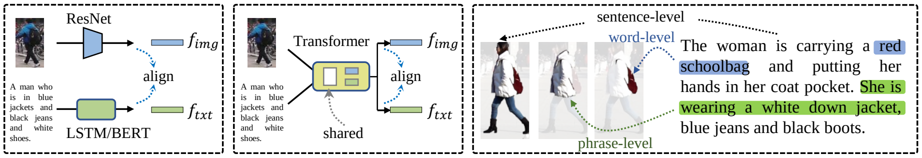
Abstract. Text-based person retrieval aims to find the query person based on a textual description. The key is to learn a common latent space mapping between visual-textual modalities. To achieve this goal, existing works employ segmentation to obtain explicitly cross-modal alignments or utilize attention to explore salient alignments. These methods have two shortcomings: 1) Labeling cross-modal alignments are time-consuming. 2) Attention methods can explore salient cross-modal alignments but may ignore some subtle and valuable pairs. To relieve these issues, we introduce an **I**mplicit **V**isual-**T**extual (**IVT**) framework for text-based person retrieval. Different from previous models, IVT utilizes a single network to learn representation for both modalities, which contributes to the visual-textual interaction. To explore the fine-grained alignment, we further propose two implicit semantic alignment paradigms: multi-level alignment (MLA) and bidirectional mask modeling (BMM). The MLA module explores **finer** matching at sentence, phrase, and word levels, while the BMM module aims to mine **more** semantic alignments between visual and textual modalities. Extensive experiments are carried out to evaluate the proposed IVT on public datasets, *i.e.,* CUHK-PEDES, RST-PReID, and ICFG-PEDES. Even without explicit body part alignment, our approach still achieves state-of-the-art performance. Code is available at: https://github.com/TencentYoutuResearch/PersonRetrieval-IVT.

Keywords: Text-based Person Retrieval, Person Search by Language, Cross-Model Retrieval

1 Introduction

Person re-identification (re-ID) has many applications, *e.g.,* finding suspects or lost children in surveillance, and tracking customers in supermarkets. As a subtask of person re-ID, text-based person retrieval (TPR) has attracted remarkable attention in recent years [23,48,39]. This is due to the fact that textual descriptions are easily accessible and can describe more details in a natural way. For

[⋆] Equal contribution [†]Corresponding Author


ResNet
f_{img}
align
A man who is in blue jackets and black jeans and white shoes.
LSTM/BERT
f_{txt}
Transformer
f_{img}
align
A man who is in blue jackets and black jeans and white shoes.
shared
f_{txt}
sentence-level
word-level
The woman is carrying a red schoolbag and putting her hands in her coat pocket. She is wearing a white down jacket, blue jeans and black boots.
phrase-level

(a) Previous separate networks (b) Our unified network (c) Implicit alignment at sentence, phrase, and word levels.

Fig. 1: **Illustration of the key idea of IVT.** Previous methods use separate models to extract features, while we utilize a single network for both modalities. The shared parameters contribute to learning a common space mapping. Besides, we explore semantic alignment using three-level matchings. Not only see finer, but also see more semantic alignments.

example, police officers usually access surveillance videos and take the deposition from witnesses. Textual descriptions can provide complementary information and even are critical in scenes where images are missing.

Text-based person retrieval needs to process visual and textual modalities, and its core is to learn a common latent space mapping between them. To achieve this goal, current works [42,16] firstly utilize different models to extract features, *i.e.,* ResNet50 [14] for visual modality, and LSTM [46] or BERT [17] for textual modality. Then they are devoted to exploring visual-textual part pairs for semantic alignment. However, these methods have at least two drawbacks that may lead to suboptimal cross-modal matching. **First**, separate models lack modality interaction. Each model usually contains many layers with a large number of parameters, and it is difficult to achieve full interaction just using a matching loss at the end. To relieve this issue, some works [45,21] on general image-text pre-training use cross-attention to conduct interaction. However, they require to encode all possible image-text pairs to compute similarity scores, which leads to quadratic time complexity at the inference stage. How to design a more suitable network for the TPR task still needs profound thinking. **Second**, labeling visual-textual part pairs, *e.g.,* head, upper body, and lower body, is time-consuming, and some pairs may be missing due to the variability of textual descriptions. For example, some text contains descriptions of hairstyles and pants, but others do not contain this information. Some researchers begin to explore implicit local alignment to mine part matching [39,48]. To ensure reliability, partial local matchings with high confidence are usually selected. However, these parts usually belong to salient regions that can be easily mined by global alignment, *i.e.,* they do not bring additional information gain. According to our observation, local semantic matching should not only see **finer**, but also see **more**. Some subtle visual-textual cues, *e.g.,* hairstyle and logo on clothes, maybe easily ignored but could be complementary to global matching.

To solve the above problems, we first introduce an **I**mplicit **V**isual-**T**extual (**IVT**) framework, which can learn representation for both modalities only using a single network (See Fig. 1(b)). This benefits from the merit that Transformer can operate on any modality that can be tokenized. To avoid the shortcomings of separate models and cross-attention, *i.e.,* separate models lack modality interaction and cross-attention models are quite slow at the inference stage, IVT supports separate feature extracting to ensure retrieval speed and shares some parameters that contribute to learning a common latent space mapping. To explore fine-grained modality matching, we further propose two implicit semantic alignment paradigms: multi-level alignment (MLA) and bidirectional mask modeling (BMM). The two paradigms do not require extra manual labeling and can be easily implemented. Specifically, as shown in Fig. 1(c), MLA aims to explore fine-grained alignment by using sentence, phrase, and word-level matchings. BMM shares a similar idea with MAE [13] and BEIT [2] in that both learn better representation through random masking. The difference is that the latter two aim at single-modal autoencoding-style reconstruction, while BMM does not reconstruct images but focuses on learning cross-modal matching. By masking a certain percentage of visual and textual tokens, BMM forces the model to mine more useful matching cues. The proposed two paradigms could not only see finer but also see more semantic alignments. Extensive experiments demonstrate the effectiveness on the TPR task.

Our contributions can be summarized as three folds: (1) We propose to tackle the modality alignment from the perspective of backbone network and introduce an Implicit Visual-Textual (IVT) framework. This is the first unified framework for text-based person retrieval. (2) We propose two implicit semantic alignment paradigms, *i.e.,* MLA and BMM, which enable the model to mine finer and more semantic alignments. (3) We conduct extensive experiments and analyses on three public datasets. Experimental results show that our approach achieves state-of-the-art performance.

2 Related Work

Text-based Person Retrieval. Considering the great potential economic and social value of text-based person retrieval, Li et al. propose the first benchmark dataset CUHK-PEDES [23] in 2017, and also build a baseline, *i.e.,* GNA-RNN, based on LSTM network. Early works utilize ResNet50 and LSTM to learn representations for visual-textual modalities, and then utilize matching loss to align them. For example, CMPM [47] associates the representations across different modalities using KL divergence. Besides aligning the features, some works study the identity cue [22], which helps learn discriminative representations. Since text-based person retrieval requires fine-grained recognition of human bodies, later works start to explore global and local associations. Some works [5,4] utilize visual-textual similarity to mine part alignments. ViTAA [42] segments the human body and utilizes k-reciprocal sampling to associate the visual and textual attributes. Surbhi et al. [1] propose to create semantic-preserving embeddings through attribute prediction. Since visual and textual attributes require pre-processing, more works attempt to use attention mechanisms to explore fine-grained alignment. PMA [16] proposes a pose-guided multi-granularity attention network. HGAN [48] splits the images into multi-scale stripes and utilizes attention to select top-k part alignments. Other works include adversarial learning and relation modeling. TIMAM [30] learns modality-invariant feature representations using adversarial and cross-modal matching objectives. A-GANet [26] introduces the textual and visual scene graphs consisting of object properties and relationships. In summary, most current works learn modality alignment by exploiting local alignments. In this work, we study the modality alignment from different perspectives, in particular, how to obtain full modality interaction and how to achieve local alignment simply. The proposed framework can effectively address these issues and achieve satisfying performance.

Transformer and Image-Text Pre-Training. Transformer [38] is firstly proposed for machine translation in the natural language processing (NLP) community. After that, many follow-up works are proposed and set new state-of-the-art one after another, such as BERT [17], GPT series [34,35,3]. The research on Transformer-based representations in computer vision is also becoming a hot spot. Early works like ViT [9] and Swin Transformer [27] take the dividing patches as input, like the discrete tokens in NLP, and achieve state-of-the-art performance on many downstream tasks. Benefiting from the merit that Transformer can operate on any modality that can be tokenized, it has been utilized in the multi- or cross-modal tasks intuitively [41]. Lu et al. [29] propose the ViLBERT to process both visual and textual inputs in separate streams that interact through co-attentional Transformer layers. Oscar [24] and VinVL [45] take the image, text, and category tags as inputs and find that the category information and stronger object detector can bring better results. Many recent works study which architecture is better for multi- or cross-modal tasks, *e.g.,* UNITER [6], ALBEF [21], and ViLT [18]. These works generally employ several training objectives to support multiple downstream vision-language tasks. The most relevant downstream task for us is image-text retrieval. To obtain modality interaction, these works generally employ cross-attention in the fusion blocks. However, they have a very slow retrieval speed at the inference stage because they need to predict the similarity of all possible image-text pairs. The ALIGN [15] and CLIP [33] are large-scale vision-language models pre-trained using only contrastive matching. These separate models are suitable for image-text retrieval, but they generally achieve satisfying performance in zero-shot settings. Some works, *e.g.,* switch Transformers [10], VLMO [40], attempt to optimize the network structure so that both retrieval and other visual-language tasks can be supported. This paper is greatly inspired by these works and aims to relieve the modality alignment for text-based person retrieval.

3 METHODOLOGY

3.1 Overview

To tackle the modality alignment in text-based person retrieval, we propose an Implicit Visual-Textual (IVT) framework as shown in Fig. 2. It consists of a unified visual-textual network and two implicit semantic alignment paradigms, *i.e.,* multi-level alignment (MLA) and bidirectional mask modeling (BMM). One key idea of IVT lies in tackling the modality alignment using a unified network. By sharing some modules, *e.g.,* the layer normalization and multi-head attention,

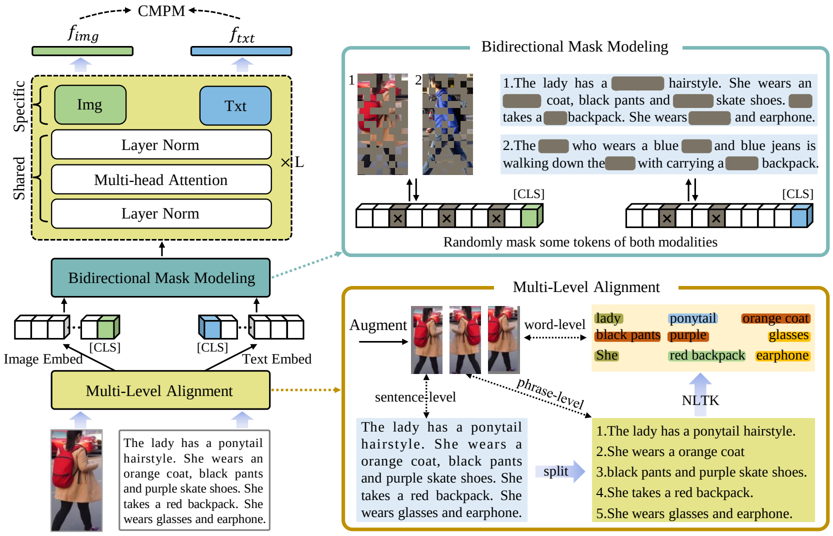

CMPM
f_{img}
f_{txt}
Specific
Shared
Img
Txt
Layer Norm
Multi-head Attention
Layer Norm
×L
Bidirectional Mask Modeling
Image Embed
[CLS]
[CLS]
Text Embed
Multi-Level Alignment
The lady has a ponytail hairstyle. She wears an orange coat, black pants and purple skate shoes. She takes a red backpack. She wears glasses and earphone.
Bidirectional Mask Modeling
1
2
1.The lady has a hairstyle. She wears an coat, black pants and skate shoes. takes a backpack. She wears and earphone.
2.The who wears a blue and blue jeans is walking down the with carrying a backpack.
[CLS]
[CLS]
Randomly mask some tokens of both modalities
Multi-Level Alignment
Augment
word-level
lady
ponytail
orange coat
black pants
purple
glasses
She
red backpack
earphone
sentence-level
phrase-level
NLTK
The lady has a ponytail hairstyle. She wears a orange coat, black pants and purple skate shoes. She takes a red backpack. She wears glasses and earphone.
split
1.The lady has a ponytail hairstyle.
2.She wears a orange coat
3.black pants and purple skate shoes.
4.She takes a red backpack.
5.She wears glasses and earphone.

Fig. 2: **Architecture of the proposed IVT Framework.** It consists of a unified visual-textual network and two implicit semantic alignment paradigms, *i.e.,* multi-level alignment (MLA) and bidirectional mask modeling (BMM). The unified visual-textual network contains parameter-shared and specific modules, which contribute to learning common space mapping. MLA aims to see "finer" by exploring local and global alignments from three-level matchings. BMM aims to see "more" by mining more semantic alignments from random masking for both modalities.

the unified network contributes to learning common space mapping between visual and textual modalities. It can also learn modality-specific cues using different modules. The two implicit semantic alignment paradigms, *i.e.,* MLA and BMM, are proposed to explore fine-grained alignment. Different from previous methods that use manually processed parts or select salient parts from attention, the two paradigms could mine not only finer but also more semantic alignments, which is another key idea of our proposed IVT.

3.2 Unified Visual-Textual Network

Embedding. As shown in Fig. 2, the input are image-text pairs, which provide the appearance characteristics of a person from visual and textual modalities. Let the image-text pairs denoted as $\{x_i, t_i, y_i\}|_{i=1}^{N}$, where x_i, t_i, y_i denote the image, text, and identity label, respectively. N is the total number of samples. For an input image $x_i \in \mathbb{R}^{H\times W\times C}$, it is firstly split into $K = H \cdot W/P^2$ patches, where P denotes the patch size, and then linearly projected into patch embeddings $\{f_k^v\}|_{k=1}^{K}$. This operation can be realized using a single convolutional layer. The patch embeddings are then prepended with a learnable class token f_{cls}^v, and added with a learnable position embedding f_{pos}^v and a type embedding f_{type}^v.

$$\mathbf{f}^v = [f_{cls}^v, f_1^v, ..., f_K^v] + f_{pos}^v + f_{type}^v. \tag{1}$$

For the input text t_i, it generally consists of one or several sentences and each sentence has a sequence of words. The pretrained word embedding is leveraged to project words into token vectors:

$$\mathbf{f}^t = [f^t_{cls}, f^t_1, ..., f^t_M, f^t_{sep}] + f^t_{pos} + f^t_{type}, \tag{2}$$

where f^t_{cls} and f^t_{sep} denote the start and end tokens. M indicates the length of tokenized subword units. f^t_{pos} is the position embedding and f^t_{type} is the type embedding.

Visual-Textual Encoder. Current works on the TPR task utilize separate models, which lack full modality interaction. Some recent works on general image-text pre-training attempt to utilize cross-attention to achieve modality interaction. However, cross-attention requires encoding all possible image-text pairs at the inference stage, which leads to a very slow retrieval speed. Based on these observations, we propose to take the unified visual-textual network for the TPR task. The unified network has a quick retrieval speed and supports modality interaction.

As shown in Fig. 2, the network follows a standard architecture of ViT [9] and stacks L blocks in total. In each block, two modalities share layer normalization (LN) and multi-head self-attention (MSA), which contribute to learning common space mapping between visual and textual modalities. It is because the shared parameters help learn common data statistics. For example, LN would calculate the mean and standard deviation of input token embeddings, and shared LN would learn the statistically common values of both modalities. This can be regarded as a "modality interaction" from a data-level perspective. As the visual and textual modalities are not the same, each block has modality-specific feed-forward layers, *i.e.,* the "Img" and "Txt" modules in Fig. 2. They are used to capture modality-specific information by switching to visual or textual inputs. The complete processing in each block can be denoted as follows:

$$\mathbf{f}^{v/t}_i = \mathrm{MSA}(\mathrm{LN}(\mathbf{f}^{v/t}_{i-1})) + \mathbf{f}^{v/t}_{i-1}, \tag{3}$$

$$\mathbf{f}^{v/t}_i = \mathrm{MLP}_{\mathrm{img/txt}}(\mathrm{LN}(\mathbf{f}^{v/t}_i)) + \mathbf{f}^{v/t}_i, \tag{4}$$

where LN denotes the layer normalization and MSA denotes the multi-head attention. i is the index of the blocks. $\mathbf{f}^{v/t}_{i-1}$ is the visual or textual output of the $(i-1)^{th}$ block and also the input of the i^{th} block. $\mathrm{MLP}_{\mathrm{img/txt}}$ denotes the modality-specific feed-forward layers. $\mathbf{f}^{v/t}_i$ is the output of the i^{th} block.

Output. The class token of the last block serves as the global representation, *i.e.,* f_{img} and f_{txt} in Fig. 2. The dimension of both feature vectors are 768, and the outputs are normalized using the LN layer.

3.3 Implicit Semantic Alignment

Multi-Level Alignment Fine-grained alignment has been demonstrated to be the key to achieving performance improvement, such as segmented attributes [42] and stripe-based parts [11,48]. These methods can be regarded as explicit part alignment, namely telling the model which visual-textual parts should be aligned. In this work, we propose an implicit alignment method, *i.e.,* multi-level alignment, which is intuitive but very effective.

As shown in Fig. 2, the input image is firstly augmented to get three types of augmented ones, *e.g.,* horizontal flipping and random cropping. The input text generally consists of one or several sentences. We split them into more short sentences according to periods and commas. These short sentences are regarded as "phrase-level" representation, which describe partial appearance characteristics of the human body. To mine finer parts, we further utilize the Natural Language Toolkit (NLTK) [28] to extract nouns and adjectives, which describe specific local characteristics, *e.g.,* bag, clothes, or pants. The three-level textual descriptions, *i.e.,* sentence-level, phrase-level, and word-level, correspond to the three types of augmented images. The three-level image-text pairs are randomly generated at each iteration. In this way, we construct a matching process that gradually refines from global to local, forcing the model to mine finer semantic alignments.

The major difference between our approach and previous work is that we do not explicitly define visual semantic parts, instead, automatically explore aligned visual parts guided by three-level textual descriptions. This is due to the following observations that inspire our TPR framework: 1) Previous explicit aligned methods lack inconsistency between the training and inference phases. During training, previous methods utilize an unsupervised way to explore local alignments, e.g., select the top-k salient alignments based on similarities. During the inference stage, only the global embeddings for each modality would be used, resulting in the inconsistent issue. 2) The explicit local alignment makes the training easier but makes it harder at the inference stage. Oversimpilified task design leads to worse generalization performance of the model. Even though we do not provide visual parts, but rather the full images, the model tries hard to mine local alignment at the training stage, thus achieving better performance at the inference stage due to the consistency.

Bidirectional Mask Modeling To automatically mine local alignment, recent methods [11,48] split images into stripes and utilize attention to select top-k part alignments. However, they ignore the fact that top-k part alignments are usually salient cues, which may have been mined by the global alignment. Therefore, these parts may bring limited information gains. We argue that local alignments should not only be finer, but also more diverse. Some subtle visual-textual cues may be complementary to global alignment.

As shown in Fig. 2, we propose a bidirectional mask modeling (BMM) method to mine more semantic alignments. For the image and text tokens, we randomly mask some percentage of them and then force the visual and textual outputs to keep in alignment. In general, the masked tokens correspond to specific patches of the image or words of the text. If specific patches or words are masked, the model would try to mine useful alignment cues from other patches or words. Let us take the lady in Fig. 2 as an example, if the salient words "orange coat" and "black pants" are masked, the model would pay more attention to other words, *e.g.,* ponytail hairstyle, red backpack. In this way, more subtle visual-textual alignments can be explored. At the training stage, this method makes it more difficult for the model to align image and text but helps it to mine more semantic alignments at the inference stage.

The above method shares a similar idea to Random Erasing [50], MAE [13], and BEIT [2]. However, Random Erasing masks only one region. MAE and BEIT aim at image reconstruction using an autoencoding style. The proposed BMM does not reconstruct images but focuses on cross-modal matching.

3.4 Loss Function

We utilize the commonly used cross-modal projection matching (CMPM) loss [47] to learn visual-textual alignment, which is defined as follows:

$$\mathcal{L}_{cmpm} = \frac{1}{B}\sum_{i=1}^{B}\sum_{j=1}^{B}\Big(p_{i,j}\cdot\log\frac{p_{i,j}}{q_{i,j}+\epsilon}\Big), \tag{5}$$

$$p_{i,j} = \frac{exp(f_i^T\cdot f_j)}{\sum_{k=1}^{B} exp(f_i^T\cdot f_k)}, \qquad q_{i,j} = \frac{y_{i,j}}{\sum_{k=1}^{B} y_{i,k}}, \tag{6}$$

where $p_{i,j}$ denotes the matching probability. f_i and f_j denote the global features of different modalities. B is the mini-batch size. $q_{i,j}$ denotes the normalized true matching probability. ϵ is a small number to avoid numerical problems.

The CMPM loss represents the KL divergence from distribution $\mathbf{q}$ to $\mathbf{p}$. Following previous work [47], the matching loss is computed in two directions, *i.e.,* image-to-text and text-to-image. The total loss can be denoted as follows:

$$\mathcal{L} = \mathcal{L}_{cmpm}^{t2v} + \mathcal{L}_{cmpm}^{v2t}. \tag{7}$$

4 Experiment

4.1 Experimental Setup

Datasets. We evaluate our approach on three benchmark datasets, *i.e.,* **CUHK-PEDES** [23], **RSTPReid** [51], and **ICFG-PEDES** [8]. Specifically, CUHK-PEDES [23] contains 40,206 images of 13,003 persons and 80,440 description sentences. It is splitted into a training set with 34,054 images and 68,126 description sentences, a validation set with 3,078 images and 6,158 description sentences, and a testing set with 3,074 images and 6,156 description sentences. RSTPReid [51] is collected from MSMT [44] and contains 20,505 images of 4,101 persons. Each image has two sentences and each sentence is no shorter than 23 words. More in detail, the training, validation, and testing sets have 3,701, 200, and 200 identities, respectively. ICFG-PEDES [8] is also collected from

Table 1: Comparison with SOTA methods on CUHK-PEDES.

Method	R1	R5	R10
CNN-RNN [36]	8.07	-	32.47
GNA-RNN [23]	19.05	-	53.64
PWM-ATH [5]	27.14	49.45	61.02
GLA [4]	43.58	66.93	76.26
MIA [31]	53.10	75.00	82.90
A-GANet [26]	53.14	74.03	81.95
ViTAA [42]	55.97	75.84	83.52
IMG-Net [43]	56.48	76.89	85.01
CMAAM [1]	56.68	77.18	84.86
HGAN [48]	59.00	79.49	86.60
NAFS [11]	59.94	79.86	86.70
DSSL [51]	59.98	80.41	87.56
MGEL [39]	60.27	80.01	86.74
SSAN [8]	61.37	80.15	86.73
NAFS [11]	61.50	81.19	87.51
TBPS [12]	61.65	80.98	86.78
TIPCB [7]	63.63	82.82	89.01
Baseline (Ours)	55.75	75.68	84.13
IVT (Ours)	**65.59**	**83.11**	**89.21**

Table 2: Comparison with SOTA methods on RSTPReid.

Method	R1	R5	R10
DSSL [51]	32.43	55.08	63.19
Baseline (Ours)	37.40	60.90	70.80
IVT (Ours)	**46.70**	**70.00**	**78.80**

Table 3: Comparison with SOTA methods on ICFG-PEDES.

Method	R1	R5	R10
Dual Path [49]	38.99	59.44	68.41
CMPM+CMPC [47]	43.51	65.44	74.26
MIA [31]	46.49	67.14	75.18
SCAN [20]	50.05	69.65	77.21
ViTAA [42]	50.98	68.79	75.78
SSAN [8]	54.23	72.63	79.53
Baseline (Ours)	44.43	63.50	71.00
IVT (Ours)	**56.04**	**73.60**	**80.22**

MSMT17 [44] and contains 54,522 images of 4,102 persons. Each image has one description sentence with an average of 37.2 words. The training and testing subsets contain 34,674 image-text pairs for 3,102 persons, and 19,848 image-text pairs for the remaining 1,000 persons, respectively. We also explore pre-training on four image captioning datasets: Conceptual Captions (CC) [37], SBU Captions [32], COCO [25] and Visual Genome (VG) [19] datasets. There are about 4M image-text pairs in total.

Evaluation Metric. The cumulative matching characteristic (CMC) curve is a precision curve that provides recognition precision for each rank. Following previous works, R1, R5, and R10 are reported when compared with state-of-the-art (SOTA) models. The mean average precision (mAP) is the average precision across all queries, which is also reported in ablation studies for future comparison.

Implementation Details. The proposed framework follows the standard architecture of ViT-Base[9].The patch size is set as 16 × 16 and the dimensions of both visual and textual features are 768. The input images are resized to 384 × 128. We use horizontal flipping and random cropping as data augmenting. At the pre-training stage, we utilize 64 Nvidia Tesla V100 GPUs with FP16 training. At the fine-tuning stage, we employ four V100 GPUs and set the mini-batch size as 28 per GPU. The SGD is used as the optimizer with the weight decay of 1e-4. The learning rate is initialized as 5e-3 with cosine learning rate decay.

4.2 Comparison with State-of-the-art Methods

In this section, we report our experimental results and compare with other SOTA methods on CUHK-PEDES [23], RSTPReid [51] and ICFG-PEDES [8]. Note that, the Baseline in Table 1, Table 2 and Table 3, denotes the vanilla IVT without pre-training, bidirectional mask modeling (BMM) and multi-level alignment

Table 4: Ablation results of components on CUHK-PEDES. "Base" denotes our baseline method. "Pre" is short of pre-training.

No.	Base	Pre	BMM	MLA	R1	R5	R10	mAP
1	✓				55.75	75.68	84.13	53.36
2	✓	✓			60.06	78.56	85.22	56.64
3	✓		✓		60.43	79.55	86.19	56.65
4	✓			✓	61.00	80.60	87.23	56.88
5	✓	✓	✓		62.88	81.60	87.54	59.34
6	✓	✓		✓	63.87	82.67	88.42	59.52
7	✓		✓	✓	64.00	82.72	88.95	58.99
8	✓	✓	✓	✓	**65.59**	**83.11**	**89.21**	**60.66**

Table 5: The computational efficiency terms of several methods. "Time" denotes the retrieval time for testing CUHK-PEDES on a Tesla V100 GPU.

Method	Architecture	Para (M)	Time (s)
ViLT [18]	Transformer	96.50	103,320
ALBEF [21]	Transformer	209.56	12,240
NAFS [11]	ResNet + BERT	189.00	78
SSAN [8]	ResNet + LSTM	97.86	31
TBPS [12]	ResNet + BiGRU	84.83	26
IVT	Transformer	166.45	42

(MLA) components. We employ BERT [17] to initialize the "txt" module and ImageNet pre-trained model from [9] to initialize other modules.

Results on CUHK-PEDES. As shown in Table 1, our baseline method achieves 55.75%, 75.68%, 84.13% on R1, R5, and R10, respectively. It already achieves comparable or even better performance compared with many works proposed in recent years, *e.g.,* MIA [31], ViTAA [42], CMAAM [1]. These experiments demonstrate the effectiveness of the unified visual-textual network for text-based person retrieval. In contrast, our proposed IVT obtains 65.59%, 83.11%, and 89.20% on these metrics, which are significantly better than our baseline method. Specifically, these results have improved considerably, *i.e.,* +9.84%, +7.43%, +5.07%, respectively. It should be noted that many recent SOTA algorithms have taken complex operations, *e.g.,* segmentation, attention, or adversarial learning. Even though, our approach outperforms existing SOTA algorithms, *e.g.,* DSSL [51] and NAFS [11], and can also be easily implemented. These results fully validate the effectiveness of our approach for text-based person retrieval.

Results on RSTPReid. As RSTPReid is newly released, only DSSL [51] has reported the results on it. As shown in Table 2, DSSL [51] achieves 32.43%, 55.08%, 63.19% on the R1, R5, R10, respectively. In contrast, the proposed method achieves 46.50%, 70.20% and 79.70%, which exceed the DSSL [51] by a large margin, *i.e.,* +14.27%, +14.92%, and +15.61%. It should be noted that our baseline method still exceeds DSSL, which benefits from our unified visual-textual network. Besides, our IVT outperforms the baseline with a large margin. The above experiments fully validate the advantages of our proposed modules.

Results on ICFG-PEDES. The experimental results on the ICFG-PEDES dataset are reported in Table 3. We can find that the baseline method achieves 44.43%, 63.50%, and 71.00% on the R1, R5, and R10, respectively. Meanwhile, our proposed IVT achieves 56.04%, 73.60%, and 80.22% on these metrics, which also fully validate the effectiveness of our proposed modules for the TPR task. Compared with other SOTA algorithms, *e.g.,* SSAN [8], ViTAA [42], our results are also better than them significantly. In contrast, even without complex operations to mine local alignments, IVT still achieves SOTA performance.

In summary, our IVT yields the best performance in terms of all metrics on three benchmark datasets. The superior performance is not only due to the well-designed unified visual-textual network, but also owing to the effective implicit semantic alignment paradigms. We hope our work can bring new insights to the text-based person retrieval community.

4.3 Ablation Study

To better understand the contributions of each component in our framework, we conduct a comprehensive empirical analysis in this section. Specifically, the results of different components of our framework on the CUHK-PEDES [23] dataset are shown in Table 4.

Effectiveness of Multi-Level Alignment. As shown in Table 4, the Baseline achieves 55.75%, 53.36% on R1 and mAP, respectively. After introducing the MLA module, the overall performance has been improved to 61.00% and 56.88%. The improvements up to +5.25% and +3.52%, respectively. The results demonstrate the effectiveness of our proposed MLA strategy. Further analysis shows that MLA enables the model to mine fine-grained matching through sentence, phrase, and word-level alignments, which in turn improves the visual and textual representations.

Effectiveness of Bidirectional Mask Modeling. The BMM strategy also plays an important role in our framework, as shown in Table 4. By comparing No.1 and No.3, we can find that R1 and mAP have been improved from 55.75%, 53.36% to 60.43%, 56.65%. The improvements up to +4.68%, +3.29% on R1 and mAP, respectively. The experimental results fully validate the effectiveness of the BMM strategy for text-based person retrieval.

Effectiveness of Pre-Training. Even without pre-training, IVT achieves 64% on R1 accuracy (see No.7), outperforming current SOTA methods, *e.g.,* NAFS (61.50%), TBPS (61.65%). To obtain better-generalized features, we pre-train our model using a large-scale image-text corpus. As illustrated in Table 4, we can find that the overall performance can also be improved significantly. Specifically, the R1 and mAP are improved from 55.75%, 53.36% to 60.06%, 56.64% with the pre-training (see No.1 and No.2). In addition, it can also be found that pre-training improves the final results by comparing the No.5/No.7 and No.8 in Table 4. Therefore, we can draw the conclusion that pre-training indeed brings more generalized features, which further boost the final matching accuracy.

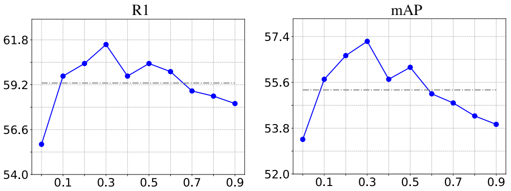


Fig. 5: **Ablation results of masking ratios on CUHK-PEDES.**

Effects of Masking Ratio. The BMM method requires setting the masking ratio. This section studies its effect on final performance. The ablation results are shown in Fig. 5. The experiments are conducted on CUHK-PEDES [23] and only the "Baseline+BMM" method is utilized. As shown in Fig. 5, the performance has been improved gradually as the masking ratio increases. When the masking ratio is set to zero, which equals the baseline method, the performance

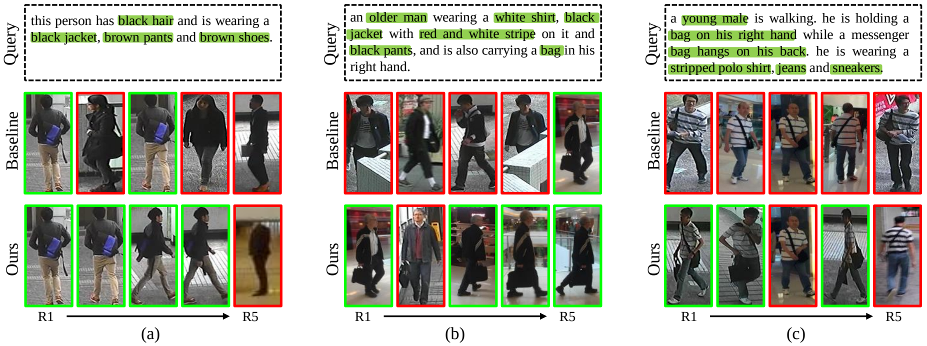

Query
this person has black hair and is wearing a black jacket, brown pants and brown shoes.
Baseline
Ours
R1
R5
(a)
Query
an older man wearing a white shirt, black jacket with red and white stripe on it and black pants, and is also carrying a bag in his right hand.
Baseline
Ours
R1
R5
(b)
Query
a young male is walking. he is holding a bag on his right hand while a messenger bag hangs on his back. he is wearing a stripped polo shirt, jeans and sneakers.
Baseline
Ours
R1
R5
(c)

Fig. 3: **The top-5 ranking results.** The green/red boxes denote the true/false results.

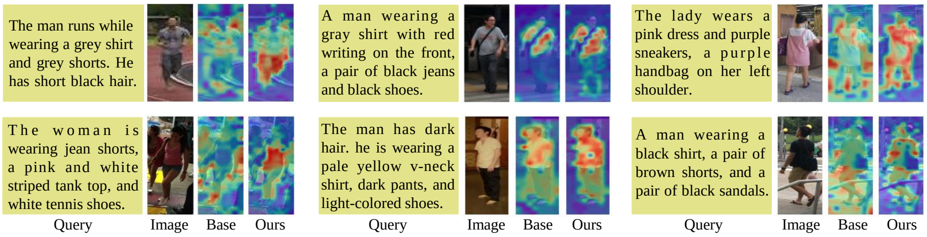

The man runs while wearing a grey shirt and grey shorts. He has short black hair.
The woman is wearing jean shorts, a pink and white striped tank top, and white tennis shoes.
A man wearing a gray shirt with red writing on the front, a pair of black jeans and black shoes.
The man has dark hair. he is wearing a pale yellow v-neck shirt, dark pants, and light-colored shoes.
The lady wears a pink dress and purple sneakers, a purple handbag on her left shoulder.
A man wearing a black shirt, a pair of brown shorts, and a pair of black sandals.
Query
Image
Base
Ours
Query
Image
Base
Ours
Query
Image
Base
Ours

Fig. 4: **Comparison of heat maps between the baseline method and IVT.**

is the worst. The R1 and mAP reach their peak values when the ratio is set as 0.3. Then as the masking ratio continues to increase, the performance gradually decreases. This is because the model cannot mine enough semantic alignments with a too large masking ratio, thus reducing the final performance.

4.4 Qualitative Results

Top-5 Ranking Results. As shown in Fig. 3, we give three examples showing the top-5 ranking results. Overall, the retrieved top-5 images show high correlations between the visual attributes and the textual descriptions, even for the false matching results. Compared with the baseline method, our proposed IVT has retrieved more positive samples. This is because it can capture more fine-grained alignments. For example, Fig. 3(b) needs to search for a person with "white shirt, black jacket with red white stripe on it and black pants". For the baseline method, the top-1 retrieved image has all these attributes, but ignores other details, *e.g.,* "older man". IVT has captured this detail and even captures the attribute "carrying a bag in his right hand". Fig. 3(c) shows a difficult case. All the retrieved images have the attributes "wearing a striped polo shirt, jeans and sneakers", but all are negative samples for the baseline method. Specifically, the baseline method ignores the description "holding a bag on his right hand" while our IVT has captured this detail and retrieved more positive samples.
Visualization of Sentence-Level Heat Map. To better understand the visual and textual alignment, we give some visualizations of sentence-level heat maps in

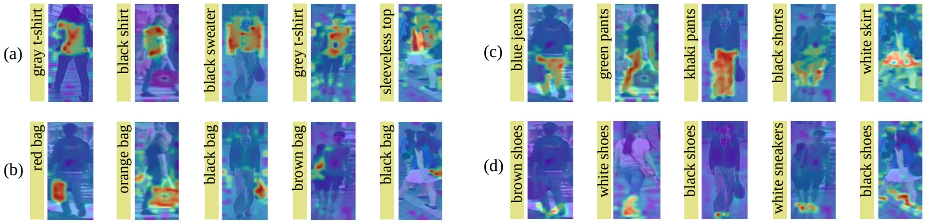


Fig. 6: **Visualization of part alignment between visual and textual modalities.** (a) Upper Body, (b) Packbag, (c) Lower Body, (d) Shoes. We compute the similarities of word-level text and all image patches. The brighter the part, the higher the similarity.

Fig. 4. The heat maps are obtained by visualizing the similarities between textual [CLS] tokens and all visual tokens. The brighter the image is, the more similar it is to the text. In general, textual descriptions can correspond to human bodies, demonstrating that the model has learned the semantic relevance of visual and textual modalities. Compared with the baseline method, IVT could focus more attributes of human bodies described by the text. For example, the man (1^{st} row, 1^{st} column) is wearing grey shorts. The baseline method has ignored the attribute, but IVT has captured it. Hence, the proposed IVT can responds to more accurate and diverse person attributes than the baseline method.

Visualization of Local Alignment. To validate the ability of fine-grained alignment, we further conduct word-level alignment experiments. As shown in Fig. 6, we give four types of human attributes, *i.e.,* upper body, packbag, lower body, and shoes. Each row in Fig. 6 shares the same attribute. The brighter the area in the image, the more similar it is to the given textual attribute descriptions. As shown in Fig. 6, our method can recognize not only salient body regions, *e.g.,* clothes and pants, but also some subtle parts, *e.g.,* handbags and shoes. These visualization results show us that our model can focus on exactly the correct body parts, given the word-level attribute description. It indicates that our approach is capable of exploring fine-grained alignments even without explicit visual-textual part alignments. This benefits from the proposed two implicit semantic alignment paradigms, *i.e.,* MLA and BMM. Therefore, we can draw the conclusion that our proposed method indeed achieves **See Finer** and **See More** for text-based person retrieval.

4.5 Computational Efficiency Analysis

In this section, we analyze the parameters and retrieval time at the inference stage. As shown in Table 5, we mainly compare recent methods in the TPR field, *e.g.,* NAFS [11], SSAN [8], TBPS [12], and typical methods in general image-text retrieval, *e.g.,* ViLT [18], ALBEF [21]. Since the parameters of LSTM/BiGRU are less than Transformer, our IVT has more parameters than SSAN and TBPS, but comparable retrieval time. Besides, the performance of these methods would be limited by the text-modeling ability of LSTM. Due to the utilization of Non-local attention, NAFS has 189M parameters and its retrieval time reaches 78s, both exceeding our IVT. Compared with general image-text retrieval methods, *e.g.,* ViLT and ALBEF, our IVT has a significant advantage in retrieval time. Specifically, ViLT needs 103,320 seconds to test CUHK-PEDES, but our IVT only needs 42 seconds. This is because they need to encode all possible image-text pairs, other than just extracting features only once. Overall, our method is competitive enough in terms of both parameters and retrieval efficiency.

5 Discussion

From the badcases in Fig. 3, we find two characteristics for the TPR task. First, the text description is usually not comprehensive, which corresponds to only part of the visual features. Second, the textual representation tends to ignore subtle features, especially for a relatively long description. By conducting extensive experiments, we got two valuable conclusions: 1) Unified network is effective for the TPR task. It maybe regarded as the backbone network in the future. 2) More subtle part alignments should be mined, other than only the salient part pairs. Even without complex operations, the proposed approaches can still mine fine-grained semantic alignments and achieve satisfying performance. We hope they can bring new insights to the TPR community.

6 Conclusion

This paper proposes to tackle the modality alignment from two perspectives: backbone network and implicit semantic alignment. First, an Implicit Visual-Textual (IVT) framework is introduced for text-based person retrieval. It can learn visual and textual representations using a single network. Benefiting from the architecture, *i.e.,* shared and specific modules, it is possible to guarantee both the retrieval speed and modality interaction. Second, two implicit semantic alignment paradigms, *i.e.,* BMM and MLA, are proposed to explore fine-grained alignment. The two paradigms could see “finer” using three-level matchings and see “more” by mining more semantic alignments. Extensive experimental results on three public datasets have demonstrated the effectiveness of our proposed IVT framework on text-based person retrieval.

7 Broader Impact

Text-based person retrieval has many potential applications in surveillance, *e.g.,* finding suspects, lost children, or elderly people. This technology can enhance the safety of the cities we live in. This work demonstrates the effectiveness of unified network and implicit alignments for the TPR task. The potential negative impact lies in that surveillance data about pedestrians may cause privacy breaches. Hence, the data collection process should be consented to by the pedestrian and the data utilization should be regulated.

Acknowledgement. This work is supported by National Natural Science Foundation of China (NO. 62102205).

References

1. 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 (WACV). pp. 2617–2625 (2020)
2. Bao, H., Dong, L., Wei, F.: BEit: BERT pre-training of image transformers. In: International Conference on Learning Representations (ICLR) (2022)
3. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. In: Advances in neural information processing systems (NeurIPS). vol. 33, pp. 1877–1901 (2020)
4. Chen, D., Li, H., Liu, X., Shen, Y., Shao, J., Yuan, Z., Wang, X.: Improving deep visual representation for person re-identification by global and local image-language association. In: Proceedings of the European conference on computer vision (ECCV). pp. 54–70 (2018)
5. 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 (2018)
6. Chen, Y.C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., Liu, J.: Uniter: Universal image-text representation learning. In: Proceedings of the European conference on computer vision (ECCV). pp. 104–120 (2020)
7. 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 **494**, 171–181 (2022)
8. 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)
9. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (ICLR) (2020)
10. Fedus, W., Zoph, B., Shazeer, N.: Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. arXiv preprint arXiv:2101.03961 (2021)
11. Gao, C., Cai, G., Jiang, X., Zheng, F., Zhang, J., Gong, Y., Peng, P., Guo, X., Sun, X.: Contextual non-local alignment over full-scale representation for text-based person search. arXiv preprint arXiv:2101.03036 (2021)
12. Han, X., He, S., Zhang, L., Xiang, T.: Text-based person search with limited data. In: The British Machine Vision Conference (BMVC) (2021)
13. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR) (2022)
14. 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 (CVPR). pp. 770–778 (2016)
15. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International Conference on Machine Learning (ICML). pp. 4904–4916. PMLR (2021)
16. 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 (AAAI). vol. 34, pp. 11189–11196 (2020)

17. Kenton, J.D.M.W.C., Toutanova, L.K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT). pp. 4171–4186 (2019)
18. Kim, W., Son, B., Kim, I.: Vilt: Vision-and-language transformer without convolution or region supervision. In: International Conference on Machine Learning (ICML). pp. 5583–5594 (2021)
19. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International journal of computer vision (IJCV) **123**(1), 32–73 (2017)
20. Lee, K.H., Chen, X., Hua, G., Hu, H., He, X.: Stacked cross attention for image-text matching. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 201–216 (2018)
21. Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., Hoi, S.C.H.: Align before fuse: Vision and language representation learning with momentum distillation. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 34 (2021)
22. 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 (ICCV). pp. 1890–1899 (2017)
23. 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 (CVPR). pp. 1970–1979 (2017)
24. Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., et al.: Oscar: Object-semantics aligned pre-training for vision-language tasks. In: Proceedings of the European conference on computer vision (ECCV). pp. 121–137. Springer (2020)
25. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Proceedings of the European conference on computer vision (ECCV). pp. 740–755 (2014)
26. Liu, J., Zha, Z.J., Hong, R., Wang, M., Zhang, Y.: Deep adversarial graph attention convolution network for text-based person search. In: Proceedings of the 27th ACM International Conference on Multimedia (MM). pp. 665–673 (2019)
27. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 10012–10022 (2021)
28. Loper, E., Bird, S.: Nltk: The natural language toolkit. arXiv preprint cs/0205028 (2002)
29. Lu, J., Batra, D., Parikh, D., Lee, S.: Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In: Advances in neural information processing systems (NeurIPS). vol. 32 (2019)
30. Nikolaos Sarafianos, Xiang Xu, I.A.K.: Adversarial representation learning for text-to-image matching. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 5813–5823 (2019)
31. Niu, K., Huang, Y., Ouyang, W., Wang, L.: Improving description-based person re-identification by multi-granularity image-text alignments. IEEE Transactions on Image Processing (TIP) **29**, 5542–5556 (2020)
32. Ordonez, V., Kulkarni, G., Berg, T.: Im2text: Describing images using 1 million captioned photographs. In: Advances in neural information processing systems (NeurIPS). vol. 24 (2011)

33. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning (ICML). pp. 8748–8763. PMLR (2021)
34. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: Improving language understanding by generative pre-training (2018)
35. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019)
36. 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 (CVPR). pp. 49–58 (2016)
37. Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL). pp. 2556–2565 (2018)
38. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems (NeurIPS). vol. 30 (2017)
39. Wang, C., Luo, Z., Lin, Y., Li, S.: Text-based person search via multi-granularity embedding learning. In: International Joint Conference on Artificial Intelligence (IJCAI). pp. 1068–1074 (2021)
40. Wang, W., Bao, H., Dong, L., Wei, F.: Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. arXiv preprint arXiv:2111.02358 (2021)
41. Wang, X., Chen, G., Qian, G., Gao, P., Wei, X.Y., Wang, Y., Tian, Y., Gao, W.: Large-scale multi-modal pre-trained models: A comprehensive survey (2022), https://github.com/wangxiao5791509/MultiModal_BigModels_Survey
42. Wang, Z., Fang, Z., Wang, J., Yang, Y.: Vitaa: Visual-textual attributes alignment in person search by natural language. In: Proceedings of the European conference on computer vision (ECCV). pp. 402–420 (2020)
43. Wang, Z., Zhu, A., Zheng, Z., Jin, J., Xue, Z., Hua, G.: Img-net: inner-cross-modal attentional multigranular network for description-based person re-identification. Journal of Electronic Imaging (JEI) **29**(4), 043028 (2020)
44. Wei, L., Zhang, S., Gao, W., Tian, Q.: Person transfer gan to bridge domain gap for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). pp. 79–88 (2018)
45. Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, L., Choi, Y., Gao, J.: Vinvl: Revisiting visual representations in vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5579–5588 (2021)
46. Zhang, S., Zheng, D., Hu, X., Yang, M.: Bidirectional long short-term memory networks for relation classification. In: Proceedings of the 29th Pacific Asia conference on language, information and computation (PACLIC). pp. 73–78 (2015)
47. 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)
48. Zheng, K., Liu, W., Liu, J., Zha, Z.J., Mei, T.: Hierarchical gumbel att ention network for text-based person search. In: Proceedings of the 28th ACM International Conference on Multimedia (MM). pp. 3441–3449 (2020)
49. Zheng, Z., Zheng, L., Garrett, M., Yang, Y., Xu, M., Shen, Y.D.: Dual-path convolutional image-text embeddings with instance loss. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) **16**(2), 1–23 (2020)

50. Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y.: Random erasing data augmentation. In: Proceedings of the AAAI conference on artificial intelligence (AAAI). vol. 34, pp. 13001–13008 (2020)

51. Zhu, A., Wang, Z., Li, Y., Wan, X., Jin, J., Wang, T., Hu, F., Hua, G.: Dssl: Deep surroundings-person separation learning for text-based person retrieval. In: Proceedings of the 29th ACM International Conference on Multimedia (MM). pp. 209–217 (2021)