BiLMa: Bidirectional Local-Matching for Text-based Person Re-identification

Takuro Fujii
Yokohama National University
tkr.fujii.ynu@gmail.com

Shuhei Tarashima
NTT Communications Corporation
tarashima@acm.org

Abstract

Text-based person re-identification (TBPReID) aims to retrieve person images represented by a given textual query. In this task, how to effectively align images and texts globally and locally is a crucial challenge. Recent works have obtained high performances by solving Masked Language Modeling (MLM) to align image/text parts. However, they only performed uni-directional (i.e., from image to text) local-matching, leaving room for improvement by introducing opposite-directional (i.e., from text to image) local-matching. In this work, we introduce Bidirectional Local-Matching (BiLMa) framework that jointly optimize MLM and Masked Image Modeling (MIM) in TBPReID model training. With this framework, our model is trained so as the labels of randomly masked both image and text tokens are predicted by unmasked tokens. In addition, to narrow the semantic gap between image and text in MIM, we propose Semantic MIM (SemMIM), in which the labels of masked image tokens are automatically given by a state-of-the-art human parser. Experimental results demonstrate that our BiLMa framework with SemMIM achieves state-of-the-art Rank@1 and mAP scores on three benchmarks.

1. Introduction

Text-based person re-identification (TBPReID) [11] aims to retrieve a target person from an image pool given a textual query. Since text queries are more user-friendly than image queries, TBPReID has been more and more expected to benefit various applications of surveillance and public safety. Existing literatures focus on how to align images and texts globally [23, 22] and/or locally [10, 4]. Particularly, recent works have demonstrated the importance of image-text local-matching [15, 20], and state-of-the-art (SOTA) methods [8, 12, 1] employ Masked Language Modeling (MLM) to align parts between image and text.

Note that, in these MLM-based TBPReID methods, a model is trained via predicting the labels of masked text tokens using unmasked image and text tokens as shown in the top of Figure 1. We argue that, however, these methods does not fully exploit local alignment between images and texts, because matching is performed only *uni-directionally* (*i.e.*, from image to text). Local-matching of the opposite direction (*i.e.*, from text to image) could also contribute to align semantically similar local image tokens (*i.e.*, patches) with corresponding text parts, but this research direction has not been explored in the literature.

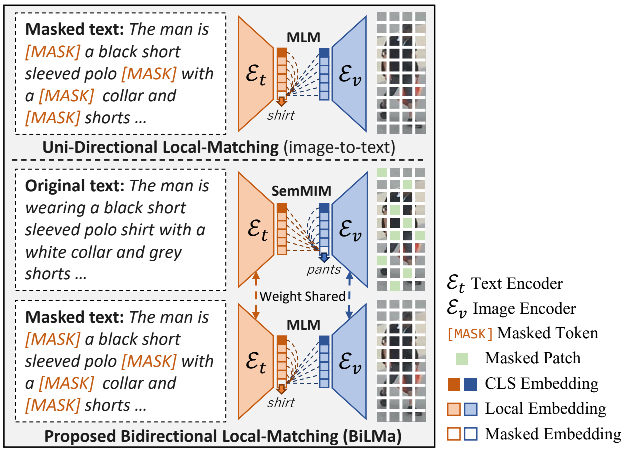
Masked text: The man is [MASK] a black short sleeved polo [MASK] with a [MASK] collar and [MASK] shorts ...
MLM
$\mathcal{E}_t$
$\mathcal{E}_v$
shirt
Uni-Directional Local-Matching (image-to-text)
Original text: The man is wearing a black short sleeved polo shirt with a white collar and grey shorts ...
SemMIM
$\mathcal{E}_t$
$\mathcal{E}_v$
pants
Weight Shared
Masked text: The man is [MASK] a black short sleeved polo [MASK] with a [MASK] collar and [MASK] shorts ...
MLM
$\mathcal{E}_t$
$\mathcal{E}_v$
shirt
Proposed Bidirectional Local-Matching (BiLMa)
$\mathcal{E}_t$ Text Encoder
$\mathcal{E}_v$ Image Encoder
[MASK] Masked Token
Masked Patch
CLS Embedding
Local Embedding
Masked Embedding

Figure 1. Overview of widely-used Uni-Directional Local-Matching and our Bidirectional Local-Matching (BiLMa). BiLMa exploits clues from both images and texts.

In this work, we propose Bidirectional Local-Matching (BiLMa) framework that can enhance local image-text alignment by jointly optimizing image-to-text MLM and text-to-image Masked Image Modeling (MIM), as illustrated in the bottom of Figure 1. In our BiLMa framework, TBPReID models are trained through predicting the labels of randomly masked both image and text tokens by all the unmasked tokens.

Notice that a straightforward approach to perform MIM in our BiLMa is to adapt existing methods [19, 2, 3, 17], which are formulated as reconstruction problems. However, we empirically found that solving reconstruction in TBPReID training is difficult (*cf.* §A.5 in the supplementary material), since it suffers huge semantic gap between modalities. To address this issue, we additionally propose a novel MIM method, named Semantic MIM (SemMIM). In our SemMIM, we formulate the MIM as a prediction of semantic labels for randomly masked image tokens by unmasked image and text tokens. With a SOTA human parser [9], we show that the semantic labels of tokens (*i.e.*, patches) can be automatically obtained.

Experimental results demonstrate that our BiLMa with SemMIM achieves SOTA Rank@1 and mAP scores on three TBPReID benchmarks. We also show that incorporating both MLM and MIM in TBPReID training (*i.e.*, BiLMa framework) leads to higher performances than the models with either MLM or MIM. To summarize, our contributions are threefold: (1) We propose Bidirectional Local-Matching (BiMLa) framework that jointly optimize MLM and MIM in TBPReID training. (2) We propose Semantic MIM (SemMIM) that can make MIM in TBPReID training tractable. (3) Experimental results demonstrate that our BiLMa with SemMIM achieves SOTA Rank@1 and mAP on three public benchmarks.

2. Related Work

Text-based Person Re-identification (TBPReID). This task was firstly introduced by [11] with a benchmark dataset. In this line of research, various solutions [7, 22, 10, 20] have been proposed, sparked by progress in the Vision-and-Language field. Particularly, recent works have achieved state-of-the-art performances by introducing Masked Language Modeling (MLM). PLIP [12] predicts masked textual tokens by masked textual tokens and visual tokens to construct the correlation between images and texts. IRRA [8] predicts masked tokens by the rest of unmasked textual tokens and visual tokens to align image and text contextualized representations and to model local dependencies. However, their MLM methods perform only uni-directional local-matching (*i.e.*, from image to text), leaving room for improvement by introducing opposite-directional (*i.e.*, from text to image) local-matching. In our work, we implement bidirectional local-matching to locally align images and texts more strongly by jointly optimizing image-to-text MLM and text-to-image MIM.

Masked Image Modeling (MIM). MIM is originally designed for self-supervised visual learning. There are various MIM strategies [3, 19, 2], all of which are formulated as reconstruction problems of randomly masked visual tokens (*i.e.*, patches) by unmasked tokens. However, we empirically found that image reconstruction from texts is difficult and not effective in TBPReID due to huge semantic gap between modalities (*cf.* §A.5). In our work, we design a novel MIM strategy, named Semantic MIM (SemMIM), for TBPReID to predict semantic labels of masked patches with textual and visual tokens.

3. Method

Our BiLMa framework can be easily deployed on top of any Transformer-based vision-language models. As a proof-of-concept, here we build BiLMa models based on IRRA [8], which is a SOTA TBPReID model at the time of this submission. In this section, we first introduce IRRA briefly, then detail our proposed BiLMa and SemMIM.

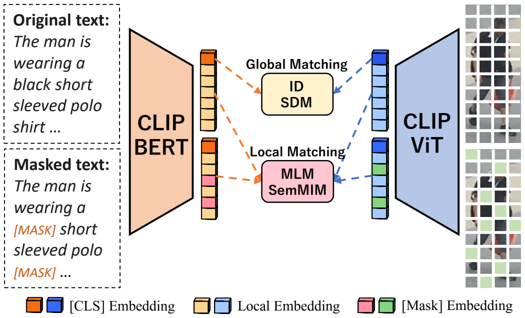

Figure 2. Overview of our BiLMa that uses ID and SDM loss for global-matching and MLM and SemMIM for local-matching.

3.1. IRRA [8]

IRRA is based on CLIP [13] image/text encoders. The image encoder takes an image I to produce a sequence of visual tokens, each of which represents a non-overlapping local token or a learnable `[CLS]` embedding. We represent the output of image encoder as $\boldsymbol{h}^V = \{\boldsymbol{h}^V_{cls}, \boldsymbol{h}^V_1, ..., \boldsymbol{h}^V_{N_v}\}$, where N_v is the number of tokens. Similarly, the text encoder takes an input text to produce a sequence of text tokens, each of which corresponds to a subword token or `[SOS]`/`[EOS]` tokens. The output of the text encoder is represented as $\boldsymbol{h}^T = \{\boldsymbol{h}^T_{sos}, \boldsymbol{h}^T_1, ..., \boldsymbol{h}^T_{N_t}, \boldsymbol{h}^T_{eos}\}$, where N_t is its token length.

IRRA employs Masked Language Modeling (MLM) to train the whole model. Specifically, during training, IRRA randomly replaces a portion of text tokens as a learnable `[MASK]` tokens. All the unmasked $\boldsymbol{h}^V$ and $\boldsymbol{h}^T$ tokens are fed into an extra encoder, which produces embeddings for correctly predicting the labels of masked tokens.

The loss function to train IRRA models is the weighted sum of SDM loss $\mathcal{L}_{sdm}$ [8] and ID loss $\mathcal{L}_{id}$ [23] for global-matching, and MLM loss $\mathcal{L}_{mlm}$ for local-matching. SDM loss is a KL-divergence between cosine similarity distributions of image-text pairs in mini-batch and true distribution, and ID loss is instance-level intra-modal matching loss. Please refer to Equation (1)-(4) in our supplementary material and the original papers [8, 23] for more details. The MLM loss is a sum of Cross-Entropy between masked textual tokens and its labels, which is defined as follows:

$$\mathcal{L}_{mlm} = -\frac{1}{|\mathcal{M}_t||\mathcal{V}|}\sum_{i\in\mathcal{M}_t}\sum_{j\in|\mathcal{V}|} y^i_j \log\frac{\exp(m^{T_m}_{i,j})}{\sum_{k=1}^{|\mathcal{V}|}\exp(m^{T_m}_{i,k})}, \quad (1)$$

where $\mathcal{M}_t$ denotes the set of masked textual tokens and $\mathcal{V}$ is the text vocabulary. y^i_j is 1 if the true label of i-th masked token is j-th vocablary in $\mathcal{V}$, and 0 otherwise. $\{m^{T_m}_{i,j}\}^{|\mathcal{V}|}_{j=1}$ is the probability of j-th word in $\mathcal{V}$ of i-th masked textual token.

3.2. Bidirectional Local-Matching (BiLMa)

BiLMa framework is illustrated in Figure 2 and 3. When we train TBPReID models with this framework, not only text tokens but also image tokens are randomly masked, then their labels are predicted by unmasked image and text tokens. More specifically, unmasked image and text tokens are fed into Cross-Modal Encoder (CME) to produce vec

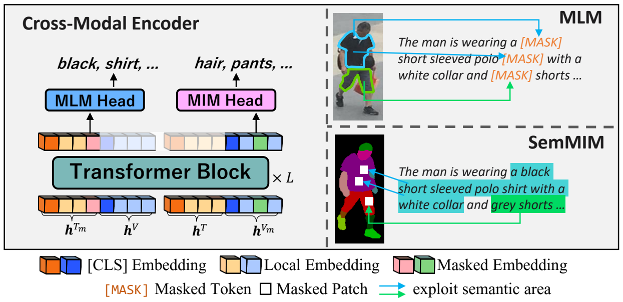

Figure 3. (Left) Cross-Model Encoder of BiLMa. (Right) MLM and our SemMIM. BiLMa enables a network to exploit visual/textual semantic area corresponding to masked textual/visual tokens via MLM/SemMIM.

Cross-Modal Encoder (CME). As shown in the left of Figure 3, CME consists of L-layer Transformer blocks, MLM head, and Masked Image Modeling (MIM) head. Given image/text encoder outputs $\boldsymbol{h}^{V/T}$ (*cf.*, §3.1), we randomly mask a portion of them to obtain masked image/text embeddings $\boldsymbol{h}^{V_m/T_m}$. Unmasked image tokens $\boldsymbol{h}^V$ and masked text tokens $\boldsymbol{h}^{T_m}$ are concatenated to be fed into the Transformer blocks, then the resulting tokens corresponding to the masked tokens are further fed into the MLM head to produce logit vectors for classification of masked words. Similarly, unmasked text tokens $\boldsymbol{h}^T$ and masked image tokens $\boldsymbol{h}^{V_m}$ are concatenated to be fed into the same Transformer blocks, then the resulting tokens corresponding to the masked tokens are further fed into the MIM head to produce logit vectors for classification of masked image tokens. Notice that we compose both MLM and MIM heads as multi-layer perceptrons of 2-layer with GELU and layer normalization. CME is removed during inference stage.

3.3. Semantic Masked Image Modeling (SemMIM)

To make text-to-image local-matching more tractable, we further propose a novel MIM method, named Semantic MIM (SemMIM). In a nutshell, given the outputs of MIM head (*i.e.*, logit vectors corresponding to the masked image tokens) and their ground truth semantic labels (*e.g.*, hair, pants as shown in the right of Figure 3), SemMIM optimize a model so as to minimize the loss of token label classification. Formally, the MIM loss $\mathcal{L}_{mim}$ is the sum of Cross-Entropy between masked image tokens and its semantic labels, which is defined as follows:

$$\mathcal{L}_{mim} = -\frac{1}{|\mathcal{M}_v||\mathcal{C}|} \sum_{i \in \mathcal{M}_V} \sum_{j \in |\mathcal{C}|} y_j^i \log \frac{\exp(m_{i,j}^{V_m})}{\sum_{k=1}^{|\mathcal{C}|} \exp(m_{i,k}^{V_m})}, \quad (2)$$

where $\mathcal{M}_v$ denotes the set of masked image tokens and $\mathcal{C}$ is the label set for tokens. y_j^i is 1 if the true label of i-th masked image token is j-th label in $\mathcal{C}$, and 0 otherwise. $\{m_{i,k}^{V_m}\}_{j=1}^{|\mathcal{C}|}$ is the probability of j-th label in $\mathcal{C}$ of i-th masked image token.

A straightforward approach to obtain such semantic labels is manual annotation, which is apparently costly and even error-prone. Therefore, we propose to introduce SOTA human parsing models to automatically give semantic labels to tokens. Specifically, given an human parser ϕ, we feed all the training images to ϕ to obtain pixel-wise semantic labels. For each token that corresponds to an image token, its semantic label is determined as the most frequent label within the token. In this work we employ a SOTA human parser [9] as ϕ. Exemplar parsing results are shown in §A.4 of our supplemental material .

This method enables to exploit the textual semantic area corresponding to masked image tokens, and make ties between them stronger. This exploitation process is illustrated in the bottom-right of Figure 3. Multi-task learning of MLM and SemMIM can achieve BiLMa, both image-to-text and text-to-image local-matching.

3.4. Loss Function

We train our model via minimizing the following loss $\mathcal{L}$:

$$\mathcal{L} = \mathcal{L}_{id} + \mathcal{L}_{sdm} + \alpha \mathcal{L}_{mlm} + \beta \mathcal{L}_{mim}. \quad (3)$$

α and β are hyperparameters to control the contribution of MLM and SemMIM, respectively.

4. Experiment

We conduct experiments on three popular benchmarks: CUHK-PEDES [11], ICFG-PEDES [5], and RSTPReid [24]. We employ widely-used Rank@K ($K = 1, 5, 10$, R@K for brevity) and mean Average Precision (mAP) as evaluation metrics, in both of which the higher is better. We compare our approach with 6 SOTA methods including ISANet [21], LBUL [18], AXM-Net [6], LGUR [14], IVT [16], CFine [20], and IRRA [8]. Due to page limitations, we leave the details of benchmarks and our implementations (including the selection of the human parser) in our supplementary material.

4.1. Comparisons with SOTA Models

The overall results on each dataset are shown in Table 1. For each dataset, we tuned the patch mask rate m_p and SemMIM loss weight β using a grid search and report the

Method	CUHK-PEDES				ICFG-PEDES				RSTPReid			
	R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP
ISANet [21]	63.92	82.15	87.69	-	57.73	75.42	81.72	-	-	-	-	-
LBUL [18]	64.04	82.66	87.22	-	-	-	-	-	45.55	68.20	77.85	-
AXM-Net [6]	64.44	80.52	86.77	58.73	-	-	-	-	-	-	-	-
LGUR [14]	65.25	83.12	89.00	-	59.20	75.32	81.56	-	-	-	-	-
IVT [16]	65.59	83.11	89.21	-	56.04	73.60	80.22	-	46.70	70.00	78.80	-
CFine [20]	69.57	85.93	91.15	-	60.83	76.55	82.42	-	50.55	72.50	81.60	-
IRRA [8]	73.38	**89.93**	**93.71**	66.13	63.46	**80.25**	**85.82**	38.06	60.20	81.30	88.20	47.17
BiLMa w/ SemMIM (Ours)	**74.03**	89.59	93.62	**66.57**	**63.83**	80.15	85.74	**38.26**	**61.20**	**81.50**	**88.80**	**48.51**

Table 1. Performance comparisons with state-of-the-art methods on CUHK-PEDES, ICFG-PEDES and RSTPReid datasets.

Components		CUHK-PEDES				ICFG-PEDES				RSTPReid			
MLM	SemMIM	R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP
		73.01	88.92	93.58	65.62	63.09	80.00	85.62	37.99	59.50	80.55	88.35	47.06
✓		73.16	89.52	**93.63**	66.00	63.60	**80.29**	85.70	38.12	59.05	80.35	87.95	46.29
	✓	73.55	89.41	93.54	66.28	63.08	80.11	85.63	37.97	59.40	80.70	87.35	46.05
✓	✓	**74.03**	**89.59**	93.62	**66.57**	**63.83**	80.15	**85.74**	**38.26**	**61.20**	**81.50**	**88.80**	**48.51**

Table 2. Ablation study on each component of BiLMa on CUHK-PEDES, ICFG-PEDES and RSTPReid datasets.

best results, while other results are described in §A.6. Following [8], the token mask rate m_t and MLM loss weight α is set to $m_t = 0.15$ and $\alpha = 1.0$.

We can clearly see that our approach (BiLMa w/ SemMIM) achieves the best Rank@1 and mAP on all the datasets. Particularly, compared to the best scores of existing methods, Rank@1 of our approach on CUHK-PEDES is 0.56% higher while mAP of ours on ICFG-PEDES is 0.37% better. On RSTPReid, ours achieves SOTA for all the metrics including Rank@1,5,10 and mAP. These results indicate the superiority and the generalization ability of our proposed approach.

4.2. Ablation Study

Next, we analyze the contribution of our proposals. Table 2 shows the results of our ablative models on three datasets. From this table, we can observe that using both MLM and SemMIM (*i.e.*, BiLMa framework) tend to achieve the best performance, indicating the good compatibility of SemMIM with MLM. Notice that our SemMIM can be solely used without MLM. Interestingly, in several cases, our model with only SemMIM outperforms the model with only MLM, which implies the strong ability of SemMIM for TBPReID model training. We also observe that our SemMIM outperforms other three MIM methods, detailed in §A.5 of the supplemental material due to page limitations.

4.3. Qualitative Analysis

Figure 4 shows two top-5 retrieval results of our model (3rd row) given a textual query shown at the top. Results of 1st and 2nd rows are our ablative models comprising only MLM or SemMIM, respectively. An image with a green frame is true positive while the one with a red frame is false positive. For clarity, phrases and their corresponding tokens are made the same color. These results show that our BiLMa w/ SemMIM can retrieve correct person more correctly. One possible reason of this superiority is that BiLMa can discriminate `very large white backpack`, `boots`, `white backpack`, `tight-fitting`, and `boots` correctly.

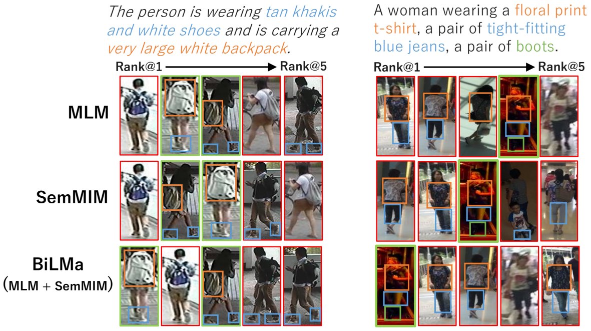
The person is wearing tan khakis and white shoes and is carrying a very large white backpack.
A woman wearing a floral print t-shirt, a pair of tight-fitting blue jeans, a pair of boots.
Rank@1
Rank@5
Rank@1
Rank@5
MLM
SemMIM
BiLMa
(MLM + SemMIM)

Figure 4. Comparison of top-5 retrieved results on CUHK-PEDES between ablative models with only MLM or SemMIM and BiLMa with both MLM and SemMIM for each text query.

5. Conclusion

In this work, we proposed Bidirectional Local-Matching (BiLMa) framework that jointly optimizes MLM and MIM in TBPReID model training. We also proposed Semantic Masked Image Modeling (SemMIM) to make text-to-image local-matching more tractable. Experiments on three TBPReID benchmarks demonstrate that our BiLMa w/ SemMIM achieves SOTA Rank@1 and mAP on all the datasets. As our future research, we plan to (1) find more helpful Masked Image/Language Modeling strategies, (2) investigate the influence of human parser's errors and consider a way to cover them.

References

[1] Yang Bai, Ming-Ming Cao, Daming Gao, Ziqiang Cao, Cheng Chen, Zhenfeng Fan, Liqiang Nie, and Min Zhang. Rasa: Relation and sensitivity aware representation learning for text-based person search. *ArXiv*, 2023.

[2] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *ArXiv*, 2021.

[3] Shuhao Cao, Peng Xu, and David A. Clifton. How to understand masked autoencoders. *ArXiv*, 2022.

[4] Yuhao Chen, Guoqing Zhang, Yujiang Lu, Zhenxing Wang, Yuhui Zheng, and Ruili Wang. Tipcb: A simple but effective part-based convolutional baseline for text-based person search. *Neurocomputing*, pages 171–181, 2021.

[5] Zefeng Ding, Changxing Ding, Zhiyin Shao, and Dacheng Tao. Semantically self-aligned network for text-to-image part-aware person re-identification. *ArXiv*, 2021.

[6] Ammarah Farooq, Muhammad Awais, Josef Kittler, and Syed Safwan Khalid. Axm-net: Implicit cross-modal feature alignment for person re-identification. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2021.

[7] Elad Hoffer and Nir Ailon. Deep metric learning using triplet network. In *International Workshop on Similarity-Based Pattern Recognition*, 2014.

[8] Ding Jiang and Mang Ye. Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2787–2797, 2023.

[9] Peike Li, Yunqiu Xu, Yunchao Wei, and Yi Yang. Self-correction for human parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[10] Shiping Li, Min Cao, and Min Zhang. Learning semantic-aligned feature representation for text-based person search. *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2724–2728, 2021.

[11] Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang. Person search with natural language description. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5187–5196, 2017.

[12] Jia li Zuo, Changqian Yu, Nong Sang, and Changxin Gao. Plip: Language-image pre-training for person representation learning. *ArXiv*, 2023.

[13] Alec Radford, Jong Wook Kim, Chris Hallacy, A. Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[14] Zhiyin Shao, Xinyu Zhang, Meng Fang, Zhi hao Lin, Jian Wang, and Changxing Ding. Learning granularity-unified representations for text-to-image person re-identification. *Proceedings of the 30th ACM International Conference on Multimedia (ACM)*, 2022.

[15] Zhiyin Shao, Xinyu Zhang, Meng Fang, Zhifeng Lin, Jian Wang, and Changxing Ding. Learning granularity-unified representations for text-to-image person re-identification. In *Proceedings of the 30th ACM International Conference on Multimedia*, page 5566–5574. Association for Computing Machinery (ACM), 2022.

[16] Xiujun Shu, Wei Wen, Haoqian Wu, Keyun Chen, Yi-Zhe Song, Ruizhi Qiao, Bohan Ren, and Xiao Wang. See finer, see more: Implicit modality alignment for text-based person retrieval. In *ECCV Workshops*, 2022.

[17] Haoqing Wang, Yehui Tang, Yunhe Wang, Jianyuan Guo, Zhi-Hong Deng, and Kai Han. Masked image modeling with local multi-scale reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2122–2131, 2023.

[18] Zijie Wang, Aichun Zhu, Jingyi Xue, Xili Wan, Chao Liu, Tian Wang, and Yifeng Li. Look before you leap: Improving text-based person retrieval by learning a consistent cross-modal common manifold. In *Proceedings of the 30th ACM International Conference on Multimedia*, page 1984–1992. Association for Computing Machinery, 2022.

[19] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: a simple framework for masked image modeling. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9643–9653, 2022.

[20] Shuanglin Yan, Neng Dong, Liyan Zhang, and Jinhui Tang. Clip-driven fine-grained text-image person re-identification. *ArXiv*, 2022.

[21] Shuanglin Yan, Hao Tang, Liyan Zhang, and Jinhui Tang. Image-specific information suppression and implicit local alignment for text-based person search. *ArXiv*, 2022.

[22] Ying Zhang and Huchuan Lu. Deep cross-modal projection learning for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[23] 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)*, pages 1 – 23, 2017.

[24] Aichun Zhu, Zijie Wang, Yifeng Li, Xili Wan, Jing Jin, Tian Wang, Fangqiang Hu, and Gang Hua. Dssl: Deep surroundings-person separation learning for text-based person retrieval. *Proceedings of the 29th ACM International Conference on Multimedia (ACM)*, 2021.