BAMG: Text-based Person Re-identification via Bottlenecks Attention and Masked Graph Modeling

Keyang Cheng[1], Wenxuan Zou[1], Hongjian Gu[2], and Anxiang Ouyang[1]

[1] Jiangsu University, Zhenjiang Xuefu Road 301, China
kycheng@ujs.edu.cn
[2] University of Science and Technology of China, Hefei Jinzhai Road 96, China

Abstract. In the realm of computer vision, traditional person re-identification (ReID) methods have primarily focused on matching pedestrian identities across varied cameras and temporal instances. Text-based Person Re-identification (TBPReID) extends these efforts by utilizing textual descriptions alongside images to enhance retrieval applications, such as tracking suspects or locating missing children. A Text-based Person Re-identification framework based on bottleneck attention and masked graph modeling(BAMG) is introduced in this paper, which incorporates the prowess of CLIP's pre-trained models into an advanced architecture. BAMG features a bottleneck fusion module for optimized modal integration, a Masked Graph Modeling (MGM) component for enhanced feature extraction, and additional supportive modules that refine the processing of multimodal data. BAMG not only enhances the alignment and interaction between text and image data but also significantly boosts the accuracy and robustness of the identification process. Through evaluations on the CUHK-PEDES dataset, the BAMG model has achieved a rank-1 accuracy of 79% and a mean average precision (mAP) of 68%. These results establish BAMG as a leading framework, setting new benchmarks for performance and adaptability in the field of multimodal learning environments focused on text-based person re-identification.

Keywords: Text-based Person Re-identification· Multimodal Learning· CLIP Models· Masked Modeling· Cross-modal Fusion

1 Introduction

Person re-identification (ReID) is a crucial task in the field of computer vision, aiming to match pedestrian identities across different cameras or time points. Traditional ReID methods have primarily focused on visual inputs. However, Text-based Person Re-identification (TBPReID), first introduced by Li et al. [25], marks a significant evolution in this domain. Unlike conventional ReID, TBPReID adopts a text-to-image retrieval approach, utilizing textual descriptions instead of photographic images to conduct pedestrian searches. This method offers considerable advantages for applications such as tracking suspects

or locating missing children, requiring robust cross-modal alignment capabilities from the models.

Historically, approaches such as those developed by [47] and others [19, 44, 30] have involved separate pre-training for each modality, followed by fine-tuning on a target dataset. However, these methods often fall short in capturing the full spectrum of intermodal correlations. More recently, multi-modal models pre-trained on datasets specifically designed for text-based pedestrian identification have shown promising results. For example, the model proposed in [18] has achieved state-of-the-art performance by leveraging these specialized datasets.

In response to the limited scale of datasets typically used in text-based pedestrian identification tasks, researchers have increasingly turned to large multi-modal pre-trained models, such as CLIP [33], which are trained on extensive datasets of hundreds of millions of samples. While these models offer considerable benefits, their direct application introduces several challenges. Firstly, there is often insufficient fusion between modalities, which compromises the adaptability of the models to specific downstream tasks. Additionally, these models frequently exhibit suboptimal occlusion resistance, struggling to maintain performance when critical visual information is obscured.

Addressing these challenges, this paper contributes the following:

- A bottleneck fusion module is proposed that enhances the integration of modal information within the CLIP model, ensuring a more seamless and effective fusion of the visual and textual modalities.
- A new approach to learning is proposed with Masked Graph Modeling (MGM) based on pedestrian structural graphs, which not only integrates pedestrian images with their corresponding structural information but also enriches the visual feature extraction process.
- The BAMG, a joint training framework, is presented, amalgamating bottleneck attention fusion, masked graph modeling, masked language learning, contrastive learning, and text-image matching learning. Comprehensive evaluations demonstrate that BAMG achieves superior results on existing text-based pedestrian datasets, setting new benchmarks for accuracy and model robustness.

2 Relative works

2.1 Text-based person re-identification

Text-based Person Re-identification (TBPReID) was first introduced by Li et al. [25], with the provision of a benchmark dataset that sparked extensive research in this domain. As the fields of vision and language have evolved, various innovative approaches have emerged, addressing complex challenges in multi-modal contexts [47, 19, 44, 30]. Recent advancements, such as the incorporation of Masked Language Modeling (MLM), have significantly improved system performance by enabling more sophisticated multimodal integrations.

For instance, PLIP [53] utilizes both masked textual and visual tokens to predict masked textual tokens, enhancing the correlation between images and texts. Similarly, IRRA [18] employs unmasked textual and visual tokens to predict masked tokens, which aids in achieving contextual alignment and capturing the nuanced dependencies between modalities. Additionally, APTM [43] extends MLM applications by predicting both masked texts and attributes, showcasing the flexibility of MLM in handling diverse data types.

Despite these innovative approaches, a critical gap remains: the majority of existing MLM methodologies focus predominantly on image-to-text matching scenarios, often neglecting text-to-image matching. This oversight limits the potential for comprehensive understanding and integration across modalities, which is crucial for tasks where visual cues are supposed to complement or even replace textual descriptions. The prevailing focus on one direction of matching underscores a significant imbalance, potentially hindering broader applications in more dynamic or varied multimodal contexts.

2.2 Visual-Language Models

Visual-Language Pretraining (VLP) has significantly improved the handling of image-language tasks by utilizing training on image-language pairings, thus offering superior generalization capabilities compared to traditional supervised pretraining methods, such as those on ImageNet. Systems like CLIP [33] and ALIGN [17], which utilize dual encoders for images and texts, have optimized performance through a bidirectional InfoNCE loss between their outputs. ALBEF [23] has focused on pre-fusion alignment of images and texts to achieve a more coherent integration of these modalities.

Inspired by natural language processing breakthroughs, visual domain adaptations such as prompts and adapter-based tuning have also gained prominence. Techniques such as learnable prompts, introduced in CoOp [17], and context vectors in CoCoOp [50], highlight the adaptability of models to dynamic contextual demands, which is crucial for tasks involving complex multimodal interactions.

The exploration of CLIP in various downstream tasks, exemplified by DenseCLIP [34] for pixel-wise predictions in segmentation and ViLD [10] for object detection, underscores the untapped versatility of CLIP in handling detailed visual tasks. The adaptability of CLIP to complex retrieval challenges, such as those demonstrated by EI-CLIP [28] and CLIP4CirDemo [3], suggests that further refinements in modal fusion techniques are necessary. In this context, the concept of Attention Bottlenecks, first introduced in [29] for integrating video and audio information through bottleneck tokens, offers a promising approach to enhancing the fusion between different modalities. This method, by facilitating deeper integration and contextual understanding between disparate data sources, could be particularly effective in improving the performance of systems like CLIP in more complex multimodal scenarios, where existing approaches have yet to fully capitalize on the potential of integrated modalities.

2.3 Masked Image Modeling

Masked Image Modeling (MIM) represents a significant advance in self-supervised learning, training models to reconstruct masked areas of images, thus reducing data redundancy and enhancing feature learning. Initial approaches like BEiT [12, 39] and MAE [12] have shown promise by learning to predict encodings of image blocks or pixels in large masked areas, respectively. SimMIM [41], with its simplified decoder structure, extends this concept by predicting pixel information across both visible and masked tokens.

However, these techniques, while innovative, have limitations, particularly in application-specific tasks such as text-based person re-identification. In their study, [7] critique existing MIM methods for their generic approach and introduce a Semantic Image Mask Modeling (semMIM) method tailored for more contextual understanding in specific applications.

Inspired by these insights and the identified gaps, we have developed a new Masked Graph Modeling (MGM) approach that incorporates human body structure into the masking process. This method not only addresses the shortcomings of generic MIM methods by providing a more targeted feature learning strategy but also leverages structural information to enhance model adaptability and performance in complex recognition tasks.

3 Methods

In this section, we present our proposed BAMG framework. The overview of BAMG is shown in Figure 1. The BAMG model comprises four main components: a graph encoder, CLIP's text and image encoders, and a final multimodal fusion encoder utilizing a bottleneck transformer. The objective function of the entire model consists of four parts: Image-Text Contrastive Loss (ITC), Masked Language Modeling Loss (MLM), Masked Graph Modeling Loss (MGM), and Image-Text Matching Loss (ITM).

3.1 Attention Bottlenecks

After processing through the CLIP model, pedestrian images and their textual descriptions are converted into tokens. Specifically, images and textual descriptions are first processed by the CLIP model to token $\mathbf{z}_{\text{img}}$ and token $\mathbf{z}_{\text{text}}$. The image-text pairs from each input network are split into patches and converted into token sequences as follows:

$$\mathbf{z} = [\mathbf{z}_{\text{img}} \,||\, \mathbf{z}_{\text{text}}] \tag{1}$$

where $||$ denotes concatenation. Subsequently, the sequences are processed by their respective cross-Transformers:

$$\mathbf{z}_{\text{img}}^{l+1} = \text{Cross-Transformer}(\mathbf{z}_{\text{img}}^{l}, \mathbf{z}^{l}; \theta_{\text{img}}) \tag{2}$$

$$\mathbf{z}_{\text{text}}^{l+1} = \text{Cross-Transformer}(\mathbf{z}_{\text{text}}^{l}, \mathbf{z}^{l}; \theta_{\text{text}}) \tag{3}$$

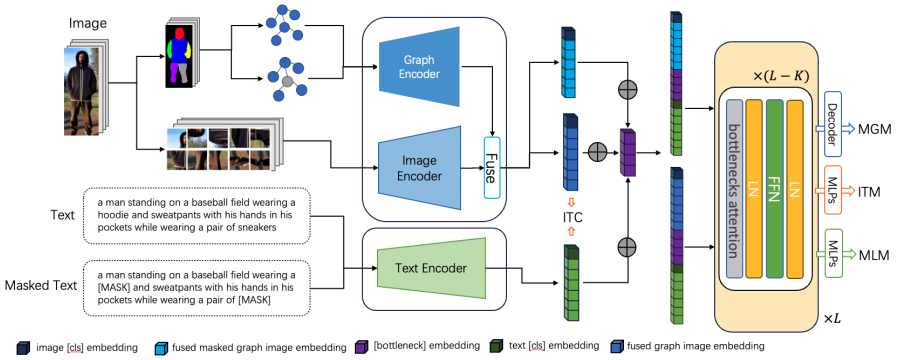

Fig. 1. Overall framework of the BAMG network.

where $\mathbf{z}^l$ represents the concatenation of $\mathbf{z}^l_{\text{img}}$ and $\mathbf{z}^l_{\text{text}}$, and θ_{img} and θ_{text} are the Transformer parameters for the respective modalities.

Specifically, these tokens are further processed through an encoder consisting of L Transformer layers, each comprising multi-head cross-attention (MCA), layer normalization (LN), and a multi-layer perceptron (MLP), as follows:

$$\mathbf{y}^l = \text{MCA}(\text{LN}(\mathbf{z}^l_1), \text{LN}(\mathbf{z}^l_2)) + \mathbf{z}^l_1 \tag{4}$$

$$\mathbf{z}^{l+1} = \text{MLP}(\text{LN}(\mathbf{y}^l)) + \mathbf{y}^l \tag{5}$$

where multi-head cross-modal attention(MCA) between two inputs is represented as $\text{MCA}(X, Y) = \text{Attention}(\mathbf{W}^Q\mathbf{X}, \mathbf{W}^K\mathbf{Y}, \mathbf{W}^V\mathbf{Y})$.

To reduce the computational complexity of the model, fusion bottleneck tokens are introduced at the input stage of the Transformer, resulting in the following input structure for the Transformer:

$$\text{z} = [\mathbf{z}_{\text{img}} \,||\, \mathbf{z}_{\text{fsn}} \,||\, \mathbf{z}_{\text{text}}] \tag{6}$$

The incorporation of bottleneck attention restricts cross-modal attention within these bottleneck tokens. For each layer, the computational process evolves as follows:

$$[\mathbf{z}^{l+1}_{\text{img}} \,||\, \hat{\mathbf{z}}^{l+1}_{\text{fsn}_1}] = \text{Transformer}([\mathbf{z}^l_{\text{img}} \,||\, \mathbf{z}^l_{\text{fsn}}]; \theta_{\text{img}}) \tag{7}$$

$$[\mathbf{z}^{l+1}_{\text{text}} \,||\, \hat{\mathbf{z}}^{l+1}_{\text{fsn}_2}] = \text{Transformer}([\mathbf{z}^l_{\text{text}} \,||\, \mathbf{z}^l_{\text{fsn}}]; \theta_{\text{text}}) \tag{8}$$

$$\mathbf{z}^{l+1}_{\text{fsn}} = \text{Avg}(\hat{\mathbf{z}}^{l+1}_{\text{fsn}_1}, \hat{\mathbf{z}}^{l+1}_{\text{fsn}_2}) \tag{9}$$

where $\mathbf{z}^{l+1}_{\text{fsn}}$ represents the updated fusion bottleneck tokens at layer l, computed as the average of the bottleneck tokens' outputs from both modalities. This method effectively alleviates the computational burden while ensuring efficient integration of cross-modal information.

By reducing the number of bottleneck tokens well below the original count, the model not only lowers its computational demands but also enhances the efficiency of information compression across modalities. This selective transfer ensures that only the most critical information is exchanged between modalities. Such strategic filtering avoids the transmission of redundant information and enhances both the efficiency and effectiveness of the model. Specifically, bottleneck tokens are updated twice to achieve information fusion: initially with image information, fo llowed by text information. The final update and optimization of information between image and text modalities are realized through the $\mathrm{Avg}(\cdot,\cdot)$ function, as depicted in Figure 2.

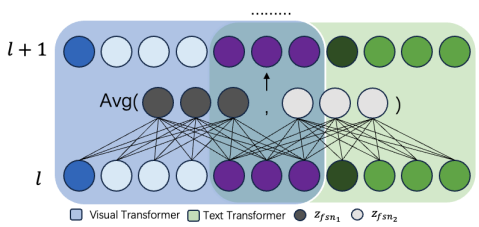

Fig. 2. Bottleneck Multimodal Fusion Mechanism.

3.2 Graph Mask Modeling Loss Based on Pedestrian Structure

Before pedestrian images are processed by the image encoder, they are first segmented using a body segmentation network, S, to differentiate various semantic parts of the body, including the head, limbs, and torso. These parts are then constructed into a graph. Subsequently, one of the nodes in this graph is masked before being input into a graph convolutional network for feature extraction. The features extracted from the graph convolutional network are then merged with the features output from the image encoder to create fused image features. These features are integrated with textual information through a multimodal fusion module, and finally, the fused image features are reconstructed back into the original pedestrian image structure. Specifically, the pedestrian image I is segmented by the segmentation network S to obtain information about different body parts:

$$\{x_{\text{head}}, x_{\text{left\_arm}}, x_{\text{right\_arm}}, x_{\text{left\_leg}}, x_{\text{right\_leg}}, x_{\text{torso}}\} = S(I) \tag{10}$$

where x_i represents the information of different body parts, the feature matrix X and the adjacency matrix A are constructed using the information from these parts and the connections between nodes. A mask operation is then performed on one of the nodes:

$$X^* = \mathrm{mask}(X, x_m) \tag{11}$$

where X^* represents the feature matrix after the mask operation, and x_m is the masked node. The masked pedestrian image is then input into a graph convolutional network to extract features, which are fused with the output features from the image encoder:

$$f_{\text{fused}} = \text{concat}(f_{\text{GNN}}(X^*, A), E_{\text{img}}(I)) \tag{12}$$

where concat denotes the concatenation and dimension reduction operation, and E_{img} denotes the image encoder operation. The fused image features are then integrated with textual features, and the original structure of the pedestrian image is restored and the MGM loss is calculated:

$$f_{\text{multi}} = \text{Bottleneck\_Transformer}(f_{\text{fused}}, f_{\text{text}}) \tag{13}$$

$$\hat{F} = D(f_{\text{multi}}) \tag{14}$$

$$\mathcal{L}_{\text{mgm}} = \frac{1}{N} \sum i = 1^N |\hat{F}_i - F_{\text{GNN},i}|^2 \tag{15}$$

where F_{GNN} denotes the graph features extracted by the Graph Neural Network (GNN).

3.3 Loss Function

Image-Text Contrastive Learning Image-Text Contrastive Learning (ITC) focuses on training the model's ability to distinguish between positive and negative pairs. For each mini-batch, we treat paired image-text (I, T) as a positive sample, and unmatched image-text pairs as negative samples. Formally, we randomly sample A pairs of images and texts in each mini-batch. Given a matched pair (I, T), its estimated matching score is as follows:

$$S_{\text{i2t}}(I) = \frac{\exp(s(F_I, F_T)/\tau)}{\sum_{i=1}^{|A|} \exp(s(F_I, F_{T^i})/\tau)}, \tag{16}$$

where τ is a learnable hyperparameter, F_I and F_T are low-dimensional features mapped from the image and attribute prompts $[CLS]$ embeddings through two different fully connected layers, respectively. $s(\cdot, \cdot)$ is the cosine similarity. The contrastive loss function is defined as:

$$\mathcal{L}_{\text{itc}} = -\frac{1}{2|A|} \sum_{(I,T)\in A} (\log S_{\text{i2t}}(I) + \log S_{\text{t2i}}(T)) \tag{17}$$

Image-Text Matching Learning The goal is to predict whether the input image and text are matched. We adopt a hard example mining strategy, selecting the highest unmatched image for each text in the mini-batch based on the similarity $S_{\text{i2t}}(T)$. The image-text pair is processed through a Cross Encoder, obtaining the matching score for the $[CLS]$ embedding:

$$p^{\text{match}}(I, T) = \text{Sigmoid}(\text{MLP}(c^{cls})) \tag{18}$$

The ITM loss is defined as:

$$\mathcal{L}_{\text{itm}} = -\frac{1}{|\bar{B}|} \sum_{(I,T)\in\bar{B}} (y^{\text{match}} \log p^{\text{match}}(I,T) \tag{19}$$

$$+ (1 - y^{\text{match}})(1 - \log p^{\text{match}}(I,T))), \tag{20}$$

when (I,T) is matched, the value of y^{match} is 1, otherwise it is 0. $\bar{B}$ is the set of all positive and negative sample pairs.

Masked Language Modeling The Masked Language Modeling (MLM) aims to restore the masked text features. For each masked text token $\hat{T}_i$, corresponding to the original text token T_i, the MLM loss is calculated as:

$$\mathcal{L}_{\text{mlm}} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j\in\mathcal{M}_i} \log P(T_{i,j}|\hat{T}_i) \tag{21}$$

where N is the number of samples, $\mathcal{M}_i$ is the set of indices of masked tokens in the ith sample and $P(T_{i,j}|\hat{T}_i)$ is the conditional probability of the jth token in the ith sample given the reconstructed text token $\hat{T}_i$.

Masked Graph Modeling Building upon the Masked Graph Modeling (MGM) approach introduced in the previous section, we define the MGM Loss to further enhance our model's ability to integrate and reconstruct visual features from fused characteristics. Let $\hat{F}_{\text{img}}$ represent the reconstructed image features from the fused characteristics, and F_{GNN} denote the graph features extracted via a Graph Neural Network (GNN). The formula for calculating the MGM Loss is as follows:

$$\mathcal{L}_{\text{mgm}} = \frac{1}{N} \sum_{i=1}^{N} |\hat{F}_{\text{img},i} - F_{\text{GNN},i}|^2 \tag{22}$$

where $|\cdot|^2$ denotes the squared norm, and N is the total number of samples. The overall objective function of the model combines the Image-Text Contrastive Loss, Image-Text Matching Loss, and the Masked Modeling Losses, defined as:

$$L = \mathcal{L}_{\text{itc}} + \mathcal{L}_{\text{itm}} + \alpha\mathcal{L}_{\text{mgm}} + \beta\mathcal{L}_{\text{mlm}} \tag{23}$$

where α and β are weight parameters that adjust the influence of each loss term.

4 Experiments

This section evaluates the model performance using the following three datasets:

CUHK-PEDES [25]: This is the first dataset introduced for Text-based Person Re-identification (TBPReID), containing 13,003 identities (IDs), 40,206 images, and 80,412 text descriptions.

ICFG-PEDES [5]: This is the second dataset specifically designed for TBP-ReID, featuring 5,452 images across 4,102 IDs, each with a text description.

RSTPReid [52]: This recently introduced TBPReID dataset includes 20,505 images from 15 cameras, covering 4,101 IDs. Each ID has five images taken by different cameras, each with two corresponding text descriptions.

4.1 Implementation Details

The experimental framework for this study was conducted on two NVIDIA RTX 4090 graphics cards. For the model setup, we utilized the pre-trained CLIP-ViT-B/16 as the image encoder, while the text encoder was the pre-trained CLIP text encoder. Each layer of the multimodal interaction encoder was configured with a hidden size of 512 and 8 attention heads. In our experiments, we applied the LoRA module to the CLIP model and froze the parameters of CLIP for fine-tuning. This approach allows us to fine-tune the model effectively while maintaining the integrity of the pre-trained weights, thus leveraging the robust feature extraction capabilities of CLIP without substantial modifications to its architecture.

All input images were resized to 384×128 pixels. The maximum length for the text token sequences was set at 77. To enhance the model's generalization ability on training data, several image augmentation techniques were employed, including random horizontal flipping, random cropping and padding, and random erasing.

The model was trained using the Adam optimizer, starting with an initial learning rate of 0.0001 and a batch size of 32. Additionally, a learning rate warm-up followed by cosine decay strategy was implemented to optimize learning rate adjustments, ensuring efficiency and effectiveness throughout the training process. The entire training spanned 45 epochs, with model saving and performance validation conducted at the end of each epoch to monitor progress and model stability.

For the loss function, the weights for different components were carefully selected based on preliminary experiments. Specifically, the weights for the masked graph modeling loss (α) and the masked language modeling loss (β) were set to 1.5 and 1.0, respectively. These values were chosen to balance the contribution of each loss component, optimizing the overall performance of the model.

4.2 Comparison with State-of-the-Art Methods

To validate the effectiveness of our proposed model, we conducted extensive comparisons with several recent advanced text-based person re-identification methods across three distinct datasets. The chosen datasets—CUHK-PEDES, ICFG-PEDES, and RSTPReid—represent varied scenarios and challenges within the

domain of text-based person re-identification, allowing us to demonstrate the robustness and adaptability of our model under different conditions. The specific comparative results are presented in Table 1.

The evaluation results show that BAMG not only competes effectively but also significantly outperforms existing state-of-the-art methods across most performance metrics. Notably, BAMG achieves higher recall rates and mean Average Precision (mAP) across all datasets, indicating its superior capability in handling complex re-identification tasks. This superior performance is primarily attributed to the innovative components integrated within the BAMG framework.

Firstly, the bottleneck fusion module significantly enhances the model's efficiency in processing multimodal data. By selectively focusing on crucial features and suppressing less relevant information, the bottleneck approach ensures that only the most pertinent features are emphasized during the fusion process. Such integration leads to a more refined and effective combination of textual and visual data, crucial for accurately matching pedestrian identities across diverse conditions. Secondly, the Masked Graph Modeling (MGM) component plays a pivotal role in capturing complex relationships within the data. By constructing and analyzing graph-based representations of the textual and visual inputs, MGM allows the model to discern subtle yet crucial attributes or features that traditional methods might overlook. The capability to effectively integrate textual and visual data proves particularly advantageous in scenarios where textual descriptions and visual appearances exhibit significant variations. This approach establishes more robust links between seemingly disparate modal data.

Table 1. Comparative experiments on three datasets.

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
Dual-Path[47]	44.4	66.26	75.07	-	38.99	59.44	68.41	-	-	-	-	-
CMPM+CMPC[44]	49.37	-	79.21	-	43.51	65.44	74.29	-	-	-	-	-
MIA[30]	53.1	75.0	82.9	-	46.49	67.14	75.18	-	-	-	-	-
SCAN[20]	-	-	-	-	50.05	69.65	77.21	-	-	-	-	-
ViTAA[40]	55.97	75.84	83.52	51.60	50.98	68.79	75.78	-	-	-	-	-
SSAN[5]	61.37	80.15	86.73	-	54.23	72.63	79.53	-	-	-	-	-
IVT[36]	65.59	83.11	89.21	-	56.04	73.6	80.22	-	46.7	70.0	78.8	-
LGUR[35]	65.25	83.12	89.00	-	59.02	75.32	81.56	-	-	-	-	-
CFine[42]	69.57	85.93	91.15	-	60.83	76.55	82.42	-	50.55	72.5	81.6	-
IRRA[18]	73.38	89.93	93.71	66.13	63.46	80.25	85.82	38.06	60.2	81.3	88.2	47.17
APTM[43]	76.53	90.04	94.15	66.91	68.51	82.99	87.56	41.22	67.50	85.70	91.45	52.56
MUM[45]	74.25	89.83	93.58	66.15	65.62	80.54	85.83	38.78	63.40	83.30	90.30	49.28
RDE[32]	75.94	90.14	94.12	67.56	67.68	82.47	87.36	40.06	65.35	83.95	89.90	50.88
Ours	79.98	92.31	94.03	68.55	71.70	86.34	89.71	42.37	69.73	87.65	93.33	55.21

4.3 Ablation Studies

To thoroughly understand the specific contributions of each component to the model's performance, we conducted a series of ablation studies. Table 2 illustrates the performance of the model under various configurations as we sequentially incorporated the bottleneck attention fusion module, the masked graph modeling module, and the masked language modeling module. The results of these ablation studies indicate that the introduction of the bottleneck attention fusion module, the masked graph modeling module, and the masked language modeling module significantly enhanced the experimental accuracy and increased the robustness of the model.

The bottleneck attention mechanism plays a pivotal role in the fusion of cross-modal information, effectively compressing and transmitting data, which in turn enhanced the robustness of the model. The integration of graph structures allowed the model to more accurately capture the structural features and complex dynamics of pedestrians. Furthermore, by incorporating masked modeling into the image feature extraction process, the model gained an improved capacity to discern and interpret the nuances of visual information, thereby boosting the accuracy of person re-identification. Additionally, masked language modeling, which involves restoring masked text features, enhanced the model's ability to understand and process textual information, further solidifying the connection between images and texts.

Table 2. Model performance under different component configurations.

Components			CUHK-PEDES				ICFG-PEDES				RSTPReid			
BF	MGM	MLM	R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP
			53.47	68.51	77.43	53.25	54.96	77.10	88.52	40.61	52.29	67.11	77.13	40.90
√			76.36	89.69	91.10	65.37	68.25	85.12	89.23	41.57	66.72	85.43	89.19	48.91
√		√	78.52	91.07	93.34	68.33	70.17	86.03	89.45	42.01	68.95	87.17	90.73	51.31
√	√	√	79.98	92.31	94.03	68.55	71.70	86.34	89.71	42.37	69.73	87.65	93.33	55.21

To comprehensively assess the impact of hyperparameters in the loss function, we designed a series of ablation studies, systematically varying the weight hyperparameters α and β for the Masked Graph Modeling Loss (MGM Loss) and the Masked Language Modeling Loss (MLM Loss). These hyperparameters play a crucial role in adjusting the model's response to different loss components and significantly affect the model's ultimate performance. The specific results of the ablation experiments are shown in Figure 3. To validate the effectiveness of the bottleneck transformer module within our model, we conducted a focused ablation study on the CUHK-PEDES dataset. This study aimes to assess the impact of incorporating the bottleneck transformer with respect to reducing the number of parameters, enhancing computational speed, and improving performance metrics. The experiment shows that the integration of the bottleneck transformer resulted in a significant parameter reduction by approximately 30%, and an increase in processing speed during inference by about 20%. More importantly,

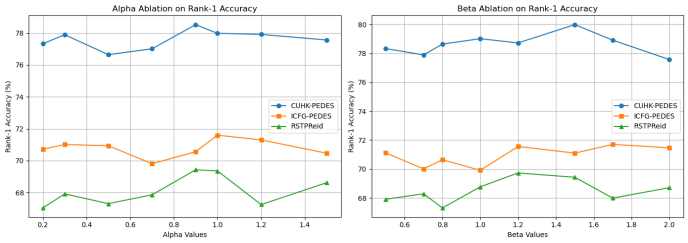

Fig. 3. Parametric Ablation Experiment

this modification led to a noticeable enhancement in the model's performance, with an approximate 3% increase in accuracy, precision, and recall across all evaluated metrics. The specific experimental results are detailed in Table 3.

Table 3. Performance Comparison of Baseline Model and Model with Bottleneck Transformer on CUHK-PEDES

Metric	Transformer	Bottleneck Transformer	Change
Execution Speed	100ms	70ms	-30%
R@1	73.93	76.36	+3.2%
R@5	85.21	89.69	+5.2%
mAP	63.19	65.37	+3.4%

4.4 Qualitative Results

In this section, we present the qualitative results comparing the performance of the BAMG model against a baseline model at Rank 5. This comparison highlights the strengths of the BAMG framework in handling complex multimodal scenarios, particularly in text-based person re-identification tasks.

Figure 4 illustrates several example cases where BAMG outperforms the baseline. These cases demonstrate BAMG's superior ability to correctly match images with textual descriptions even under challenging conditions such as variations in lighting, pose, and background clutter. Each example provides a side-by-side view of the top 5 results from both the BAMG and the baseline models, showcasing the more accurate and relevant matches produced by our model.

To gain a more intuitive understanding of the performance of our proposed BAMG model in feature extraction and cross-modal information fusion, we conducted an analysis of attention weight heatmaps. These visualizations reveal how the model processes textual and image data, shedding light on the inner workings of the model.

Figure 5 displays the attention weight distributions for several example images. These heatmaps illustrate the relationships between the image regions the model focuses on and the corresponding textual descriptions during cross-modal

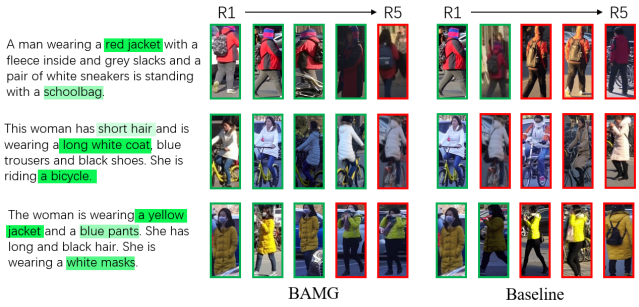

Fig. 4. Comparison of top-5 retrieved results on RSTPReid between Baseline and BAMG for each text query.

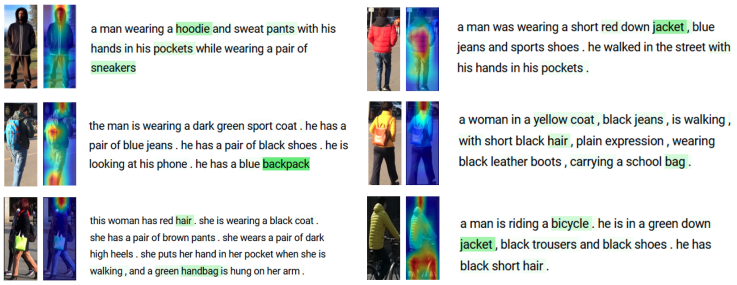

Fig. 5. Attention weight heatmaps for example images.

information fusion. It is evident that the model effectively attends to key visual areas mentioned in the text descriptions, thereby enhancing the accuracy of person re-identification.

5 Conclusion

This paper introduced the Bottleneck Attention and Masked Graph Modeling-based Text-based Person Re-identification framework (BAMG), which significantly enhanced performance across various established TBPReID datasets. By incorporating the bottleneck attention fusion module, the masked graph modeling module, and the masked language modeling module, our framework not only optimized the fusion of information across different modalities but also significantly improved the accuracy and robustness of the model in processing complex multimodal data. The effectiveness of these modules was further validated by ablation studies, underscoring their critical contributions to the overall enhancement of system performance.

Moving forward, we aim to investigate more efficient fusion strategies and advanced self-supervised learning techniques to overcome the remaining challenges in Text-based Person Re-identification. Additionally, the scalability and

adaptability of the model will be the focus of our research, aiming to accommodate a wider range of practical application scenarios and more diverse data environments.

6 Acknowledgments

This work was supported by National Natural Science Foundation of China (Grant NO.62372215) and Special fund project of Jiangsu Science and Technology Plan (Grant NO.BE2022781)

References

1. Aggarwal, S., Babu, R.V., Chakraborty, A.: Text-based person search via attribute-aided matching. 2020 IEEE Winter Conference on Applications of Computer Vision (WACV) pp. 2606–2614 (2020), https://api.semanticscholar.org/CorpusID:214603873
2. Aghajanyan, A., Zettlemoyer, L., Gupta, S.: Intrinsic dimensionality explains the effectiveness of language model fine-tuning. ArXiv **abs/2012.13255** (2020), https://api.semanticscholar.org/CorpusID:229371560
3. Baldrati, A., Bertini, M., Uricchio, T., Bimbo, A.: Composed image retrieval using contrastive learning and task-oriented clip-based features. ACM Transactions on Multimedia Computing, Communications and Applications **20**, 1 – 24 (2023), https://api.semanticscholar.org/CorpusID:261065158
4. Chen, Y., Zhang, G., Lu, Y., Wang, Z., Zheng, Y., Wang, R.: Tipcb: A simple but effective part-based convolutional baseline for text-based person search. Neurocomputing **494**, 171–181 (2021), https://api.semanticscholar.org/CorpusID:235187010
5. Ding, Z., Ding, C., Shao, Z., Tao, D.: Semantically self-aligned network for text-to-image part-aware person re-identification. ArXiv **abs/2107.12666** (2021), https://api.semanticscholar.org/CorpusID:236447478
6. Farooq, A., Awais, M., Kittler, J., Khalid, S.S.: Axm-net: Implicit cross-modal feature alignment for person re-identification. In: AAAI Conference on Artificial Intelligence (2021), https://api.semanticscholar.org/CorpusID:250294601
7. Fujii, T., Tarashima, S.: Bilma: Bidirectional local-matching for text-based person re-identification. 2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW) pp. 2778–2782 (2023), https://api.semanticscholar.org/CorpusID:261681910
8. Gao, C., Cai, G., Jiang, X., Zheng, F., Zhang, J., Gong, Y., Peng, P., Guo, X.W., Sun, X.: Contextual non-local alignment over full-scale representation for text-based person search. ArXiv **abs/2101.03036** (2021), https://api.semanticscholar.org/CorpusID:231419065
9. Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H., Qiao, Y.J.: Clip-adapter: Better vision-language models with feature adapters. ArXiv **abs/2110.04544** (2021), https://api.semanticscholar.org/CorpusID:238583492
10. Gu, X., Lin, T.Y., Kuo, W., Cui, Y.: Open-vocabulary object detection via vision and language knowledge distillation. In: International Conference on Learning Representations (2021), https://api.semanticscholar.org/CorpusID:238744187
11. Han, X., He, S., Zhang, L., Xiang, T.: Text-based person search with limited data. In: British Machine Vision Conference (2021), https://api.semanticscholar.org/CorpusID:239050116

12. He, K., Chen, X., Xie, S., Li, Y., Doll'ar, P., Girshick, R.B.: Masked autoencoders are scalable vision learners. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 15979–15988 (2021), https://api.semanticscholar.org/CorpusID:243985980
13. Hoi, S.C.H., Buntine, W.L. (eds.): Proceedings of the 4th Asian Conference on Machine Learning, ACML 2012, Singapore, Singapore, November 4-6, 2012, JMLR Proceedings, vol. 25. JMLR.org (2012), http://jmlr.org/proceedings/papers/v25/
14. Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., de Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S.: Parameter-efficient transfer learning for nlp. ArXiv **abs/1902.00751** (2019), https://api.semanticscholar.org/CorpusID:59599816
15. Hsu, C., Lee, W.S. (eds.): Proceedings of the 3rd Asian Conference on Machine Learning, ACML 2011, Taoyuan, Taiwan, November 13-15, 2011, JMLR Proceedings, vol. 20. JMLR.org (2011), http://jmlr.org/proceedings/papers/v20/
16. Hu, J.E., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Chen, W.: Lora: Low-rank adaptation of large language models. ArXiv **abs/2106.09685** (2021), https://api.semanticscholar.org/CorpusID:235458009
17. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q.V., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. ArXiv **abs/2102.05918** (2021), https://api.semanticscholar.org/CorpusID:231879586
18. Jiang, D., Ye, M.: Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 2787–2797 (2023), https://api.semanticscholar.org/CorpusID:257663606
19. Jing, Y., Si, C., Wang, J., Wang, W., Wang, L., Tan, T.: Pose-guided multi-granularity attention network for text-based person search. In: AAAI Conference on Artificial Intelligence (2018), https://api.semanticscholar.org/CorpusID:208309887
20. Lee, K.H., Chen, X., Hua, G., Hu, H., He, X.: Stacked cross attention for image-text matching. ArXiv **abs/1803.08024** (2018), https://api.semanticscholar.org/CorpusID:3994012
21. Lester, B., Al-Rfou, R., Constant, N.: The power of scale for parameter-efficient prompt tuning. In: Conference on Empirical Methods in Natural Language Processing (2021), https://api.semanticscholar.org/CorpusID:233296808
22. Li, J., Li, D., Xiong, C., Hoi, S.C.H.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning (2022), https://api.semanticscholar.org/CorpusID:246411402
23. Li, J., Selvaraju, R.R., Gotmare, A.D., Joty, S.R., Xiong, C., Hoi, S.C.H.: Align before fuse: Vision and language representation learning with momentum distillation. In: Neural Information Processing Systems (2021), https://api.semanticscholar.org/CorpusID:236034189
24. Li, S., Cao, M., Zhang, M.: Learning semantic-aligned feature representation for text-based person search. ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) pp. 2724–2728 (2021), https://api.semanticscholar.org/CorpusID:245124345
25. Li, S., Xiao, T., Li, H., Zhou, B., Yue, D., Wang, X.: Person search with natural language description. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 5187–5196 (2017), https://api.semanticscholar.org/CorpusID:515843

26. Li, X.L., Liang, P.: Prefix-tuning: Optimizing continuous prompts for generation. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) **abs/2101.00190** (2021), https://api.semanticscholar.org/CorpusID:230433941
27. Li, Y., Fan, H., Hu, R., Feichtenhofer, C., He, K.: Scaling language-image pre-training via masking. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 23390–23400 (2022), https://api.semanticscholar.org/CorpusID:254125280
28. Ma, H., Zhao, H., Lin, Z., Kale, A., Wang, Z., Yu, T., Gu, J., Choudhary, S., Xie, X.: Ei-clip: Entity-aware interventional contrastive learning for e-commerce cross-modal retrieval. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 18030–18040 (2022), https://api.semanticscholar.org/CorpusID:250445927
29. Nagrani, A., Yang, S., Arnab, A., Jansen, A., Schmid, C., Sun, C.: Attention bottlenecks for multimodal fusion. ArXiv **abs/2107.00135** (2021), https://api.semanticscholar.org/CorpusID:235694621
30. 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 **29**, 5542–5556 (2019), https://api.semanticscholar.org/CorpusID:195345251
31. Ong, C.S., Ho, T.B. (eds.): Asian Conference on Machine Learning, ACML 2013, Canberra, ACT, Australia, November 13-15, 2013, JMLR Proceedings, vol. 29. JMLR.org (2013), http://jmlr.org/proceedings/papers/v29/
32. Qin, Y., Chen, Y., Peng, D., Peng, X., Zhou, J.T., Hu, P.: Noisy-correspondence learning for text-to-image person re-identification. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 27187–27196 (2023), https://api.semanticscholar.org/CorpusID:261048736
33. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning (2021), https://api.semanticscholar.org/CorpusID:231591445
34. Rao, Y., Zhao, W., Chen, G., Tang, Y., Zhu, Z., Huang, G., Zhou, J., Lu, J.: Denseclip: Language-guided dense prediction with context-aware prompting. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 18061–18070 (2021), https://api.semanticscholar.org/CorpusID:244800733
35. Shao, Z., Zhang, X., Fang, M., hao Lin, Z., Wang, J., Ding, C.: Learning granularity-unified representations for text-to-image person re-identification. Proceedings of the 30th ACM International Conference on Multimedia (2022), https://api.semanticscholar.org/CorpusID:250627620
36. Shu, X., Wen, W., Wu, H., Chen, K., Song, Y.Z., Qiao, R., Ren, B., Wang, X.: See finer, see more: Implicit modality alignment for text-based person retrieval. In: ECCV Workshops (2022), https://api.semanticscholar.org/CorpusID:251643466
37. Sugiyama, M., Yang, Q. (eds.): Proceedings of the 2nd Asian Conference on Machine Learning, ACML 2010, Tokyo, Japan, November 8-10, 2010, JMLR Proceedings, vol. 13. JMLR.org (2010), http://jmlr.org/proceedings/papers/v13/
38. Suo, W., Sun, M., Niu, K., Gao, Y., Wang, P., Zhang, Y., Wu, Q.: A simple and robust correlation filtering method for text-based person search. In: European Conference on Computer Vision (2022), https://api.semanticscholar.org/CorpusID:253448473

39. Wang, W., Bao, H., Dong, L., Bjorck, J., Peng, Z., Liu, Q., Aggarwal, K., Mohammed, O.K., Singhal, S., Som, S., Wei, F.: Image as a foreign language: Beit pretraining for all vision and vision-language tasks. ArXiv **abs/2208.10442** (2022), https://api.semanticscholar.org/CorpusID:251719655
40. Wang, Z., Fang, Z., Wang, J., Yang, Y.: Vitaa: Visual-textual attributes alignment in person search by natural language. ArXiv **abs/2005.07327** (2020), https://api.semanticscholar.org/CorpusID:218665538
41. Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., Hu, H.: Simmim: a simple framework for masked image modeling. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 9643–9653 (2021), https://api.semanticscholar.org/CorpusID:244346275
42. Yan, S., Dong, N., Zhang, L., Tang, J.: Clip-driven fine-grained text-image person re-identification. IEEE Transactions on Image Processing **32**, 6032–6046 (2022), https://api.semanticscholar.org/CorpusID:252993001
43. Yang, S., Zhou, Y., Zheng, Z., Wang, Y., Zhu, L., Wu, Y.: Towards unified text-based person retrieval: A large-scale multi-attribute and language search benchmark. Proceedings of the 31st ACM International Conference on Multimedia (2023), https://api.semanticscholar.org/CorpusID:259075465
44. Zhang, Y., Lu, H.: Deep cross-modal projection learning for image-text matching. In: European Conference on Computer Vision (2018), https://api.semanticscholar.org/CorpusID:52957778
45. Zhao, Z., Liu, B., Lu, Y., Chu, Q., Yu, N.: Unifying multi-modal uncertainty modeling and semantic alignment for text-to-image person re-identification. In: AAAI Conference on Artificial Intelligence (2024), https://api.semanticscholar.org/CorpusID:268692441
46. Zheng, K., Liu, W., Liu, J., Zha, Z., Mei, T.: Hierarchical gumbel attention network for text-based person search. Proceedings of the 28th ACM International Conference on Multimedia (2020), https://api.semanticscholar.org/CorpusID:222278039
47. 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**, 1 – 23 (2017), https://api.semanticscholar.org/CorpusID:49867191
48. Zhou, C., Loy, C.C., Dai, B.: Extract free dense labels from clip. In: European Conference on Computer Vision (2021), https://api.semanticscholar.org/CorpusID:251105026
49. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision **130**, 2337 – 2348 (2021), https://api.semanticscholar.org/CorpusID:237386023
50. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 16795–16804 (2022), https://api.semanticscholar.org/CorpusID:247363011
51. Zhou, Z., Washio, T. (eds.): Advances in Machine Learning, First Asian Conference on Machine Learning, ACML 2009, Nanjing, China, November 2-4, 2009. Proceedings, Lecture Notes in Computer Science, vol. 5828. Springer (2009). https://doi.org/10.1007/978-3-642-05224-8, http://dx.doi.org/10.1007/978-3-642-05224-8
52. 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. Proceedings of the 29th ACM International Conference on Multimedia (2021), https://api.semanticscholar.org/CorpusID:237490866

53. li Zuo, J., Yu, C., Sang, N., Gao, C.: Plip: Language-image pre-training for person representation learning. ArXiv **abs/2305.08386** (2023), https://api.semanticscholar.org/CorpusID:258685651