Proceedings of Machine Learning Research vol 278:1–11, 2025 2025 2nd International Conference on Machine Learning and Intelligent Computing

MEFN: A Multi-scale Entropy-aware Fusion Network For Image-Text Retrieval

Jinjin Liu[*] LIUJINJIN0809@ZUT.EDU.CN
Zhongyuan University of Technology, Zhengzhou, China

Changchang Fan 360667241@QQ.COM
Zhongyuan University of Technology, Zhengzhou, China
[*]*Corresponding author*

Editors: Nianyin Zeng, Ram Bilas Pachori and Dongshu Wang

Abstract

Image-Text Retrieval(ITR), a crucial task in multi-modal learning, aims to achieve cross-modal information retrieval through semantic alignment and matching between images and text. With the advancement of deep learning, significant progress has been made in the accuracy and efficiency of ITR methods. However, existing approaches still face challenges such as modality heterogeneity, information redundancy, and insufficient multi-scale feature alignment between images and text. To address these issues, this paper proposes an Image-Text Retrieval method based on a Multi-scale Entropy-aware Fusion Network (MEFN). By introducing entropy-aware modeling and multi-scale attention mechanisms, this method enhances the correlation between image and text features, further improving cross-modal semantic matching capabilities. Specifically, MEFN first guides the fusion of image and text features through an entropy-aware model, then finely models multi-scale features using local and global attention mechanisms to generate efficient image-text fusion representations. Experimental results demonstrate that MEFN significantly improves the accuracy and robustness of image-text retrieval compared to mainstream methods on benchmark datasets such as Flickr30K and MSCOCO, especially showing superior performance in fine-grained object matching and complex scenarios. This study provides a new perspective for image-text retrieval methods and holds promise for further applications in multi-lingual image-text retrieval and video-text retrieval fields.

Keywords: Image-Text Retrieval; Multi-modal Learning; Entropy Awareness; Attention Mechanism; Cross-Modal Fusion

1. Introduction

Image-Text Retrieval (ITR), as a core task in cross-modal retrieval, aims to achieve semantic matching and efficient retrieval between visual and textual modalities. It has found widespread applications in fields such as information retrieval, intelligent recommendation, and social media analysis. In recent years, with the rapid advancement of deep learning technologies and the continuous accumulation of large-scale image-text paired data, ITR methods have made significant progress. Among them, the research paradigm represented by Vision-Language Pretraining (VLP) has become the mainstream. By constructing a unified cross-modal semantic space, VLP-based methods have greatly improved retrieval accuracy and generalization capability (Chen et al., 2023).

Nevertheless, current ITR systems still face two key challenges: modality heterogeneity and information redundancy. On the one hand, images and texts differ inherently in perceptual forms and semantic structures, making it difficult to align cross-modal features and hindering the establishment of deep semantic associations (Zhang et al., 2021). On the other hand, multimodal data often contain redundant content, especially in the form of low-value regions in images, which may introduce semantic noise and undermine the model's robustness and discriminative power in complex scenarios. To address these issues, some studies have incorporated attention mechanisms and information bottleneck theory to enhance the model's focus on key information and improve the selectivity of semantic modeling (Liang et al., 2023).

Building on this, VLP models leverage contrastive learning with large-scale image-text pairs to construct a shared cross-modal semantic space, effectively alleviating the alignment barriers caused by modality heterogeneity. Meanwhile, techniques such as masked modeling and region-level matching help reduce redundant information (Chen et al., 2023; Gan et al., 2022; Shrestha et al., 2023). However, mainstream VLP approaches typically focus on single-scale modeling and global semantic fusion, which limits their ability to capture the differences and uncertainties among multi-level and fine-grained semantics. This, in turn, constrains their performance in complex retrieval tasks.

To address these challenges, this paper proposes a novel Image-Text Retrieval method based on a Multi-scale Entropy-aware Fusion Network (MEFN). Built upon the VLP architecture, MEFN introduces an entropy-aware mechanism to quantify the information density of features and employs multi-scale attention modeling to enhance the representation of different semantic granularities. This approach significantly improves the discriminability and alignment of cross-modal features. Experimental results on benchmark datasets such as Flickr30K and MSCOCO demonstrate that MEFN substantially outperforms existing mainstream methods, validating its effectiveness and practical value in fine-grained semantic understanding and complex scene retrieval.

2. Related Work

Image-Text Retrieval (ITR) aims to establish semantic alignment between visual and textual modalities and is a vital direction in cross-modal research. With the development of deep learning and Vision-Language Pretraining (VLP) technologies, ITR methods have primarily focused on feature extraction and alignment, attention mechanisms, multimodal fusion, and compressed representation modeling.

2.1. Modeling Modality Heterogeneity and Semantic Alignment Strategies

In early approaches, image features were mainly extracted using Convolutional Neural Networks (CNNs), while textual features were modeled using RNNs or Transformer-based architectures. VLP models such as CLIP and ALIGN significantly enhanced semantic alignment precision through cross-modal contrastive learning, driving the development of a unified semantic space (Cao et al., 2022). On this basis, further research explored global alignment strategies (e.g., VSE++, SCAN) and fine-grained alignment methods (e.g., ViLBERT, ALIGN), reinforcing matching between image regions and textual fragments via attention mechanisms (Vala and Jaliya, 2022; Wang et al., 2024). However, these models still struggle with robustness and semantic discrimination when facing complex or noisy data.

Modality heterogeneity—i.e., the inherent differences between images and texts in structural representation, semantic units, and perceptual modes—is one of the core limitations affecting ITR model accuracy. To address this issue, various alignment strategies have been proposed. For example, Ji et al. Ji et al. (2022) introduced the HMGR (Heterogeneous Memory Enhanced Graph Reasoning Network), which incorporates a heterogeneous memory mechanism and graph reasoning paths to enhance deep semantic interaction between images and texts, significantly improving the model's adaptability to complex modality structures. Zhang et al. Zhang et al. (2021) proposed the HCMSL model, which jointly learns cross-modal similarity through deep networks and introduces a Siamese structure to learn intra-modal similarity, thereby mitigating inconsistencies in cross-modal semantic space structures. Additionally, Zhou et al. Zhou et al. (2023) noted in their review that multi-scale feature alignment, local semantic matching, and cross-modal graph modeling will play key roles in addressing modality heterogeneity challenges in the future.

2.2. Information Redundancy Suppression and Fusion Mechanism Optimization

Traditional multimodal fusion methods often rely on static strategies, making them ineffective at handling redundancy and semantic conflicts between modalities. To tackle this, attention mechanisms have been widely adopted to enhance the model's focus on critical regions (Vala and Jaliya, 2022), while multimodal tensor fusion and re-ranking techniques have achieved notable improvements in accuracy (Wang et al., 2019). However, these methods typically consume significant computational resources, making it difficult to balance accuracy with efficiency (Cao et al., 2022; Zhu et al., 2023).

During multimodal fusion, different modalities often contain redundant or irrelevant information—such as background regions in images or redundant descriptions in text—which can interfere with the discriminative quality of semantic representations. To improve representation quality and fusion performance, Mai et al. (2022) proposed the Multimodal Information Bottleneck (MIB) architecture. From an information-theoretic perspective, MIB maximizes the mutual information between representations and task objectives while suppressing redundant information from the input, effectively enhancing the compactness and discriminative power of multimodal representations. For high-redundancy scenarios like video, Wu et al. (2023) introduced the Denoising Bottleneck Fusion (DBF) method, which combines bottleneck mechanisms with mutual information maximization strategies to filter intra and inter modal redundancy and noise, thereby improving fusion accuracy. Additionally, Liang et al. (2023) proposed an information decomposition framework (involving redundancy, uniqueness, and synergy), inspired by human annotation behavior, to quantify the information structure in multimodal interactions, offering new perspectives and quantitative support for redundancy modeling.

2.3. Multi-scale Modeling and Proposed Method

Existing methods generally lack the capacity to distinguish and model semantics at different granularities, resulting in insufficient multi-scale alignment (Yang et al., 2024). Meanwhile, the complexity of current models increases computational overhead, posing efficiency challenges in real-world scenarios (Cao et al., 2022; Wang et al., 2025). To address this issue, this paper proposes an ITR framework based on a Multi-scale Entropy-aware Fusion Mechanism. The proposed method introduces entropy-based measurement and scale-adaptive modeling to enhance the flexibility and discriminability of cross-modal feature alignment. Combined with the VLP backbone, it achieves high performance while maintaining computational efficiency (Cao et al., 2022).

3. Proposed Method MEFN

3.1. Overall Framework

The proposed Multi-scale Entropy-aware Fusion Network (MEFN) aims to enhance cross-modal retrieval performance, particularly for semantic alignment and retrieval efficiency between images and text. The framework consists of four main modules: Feature Extraction Module, Entropy-Aware Modeling Module, Multi-scale Attention Fusion Module, and Image-Text Joint Representation and Retrieval Module. These modules work collaboratively to optimize image-text alignment, ensuring efficient and accurate retrieval across modalities.

- Feature Extraction Module: Extracts raw representations from images and text. Images are processed through visual encoders (e.g., CNN or ViT) to obtain spatial structure information, while text is encoded into context-sensitive word embeddings using models like BERT (Zhang et al., 2020).

- Entropy-Aware Modeling Module: Introduces the Modality Fusion Entropy (MFE) mechanism to measure the information coupling strength between image and text features, thereby optimizing subsequent feature weighting.

- Multi-scale Attention Fusion Module: Employs local attention mechanisms to capture fine-grained semantic alignment (e.g., matching image regions with text words) and global attention mechanisms to focus on overall semantic consistency. These two attention mechanisms work together to enhance cross-modal semantic fusion.

- Image-Text Joint Representation and Retrieval Module: Represents fused image-text features in a unified embedding space, optimizes alignment using a triplet loss training strategy, and finally performs image-text matching and retrieval.

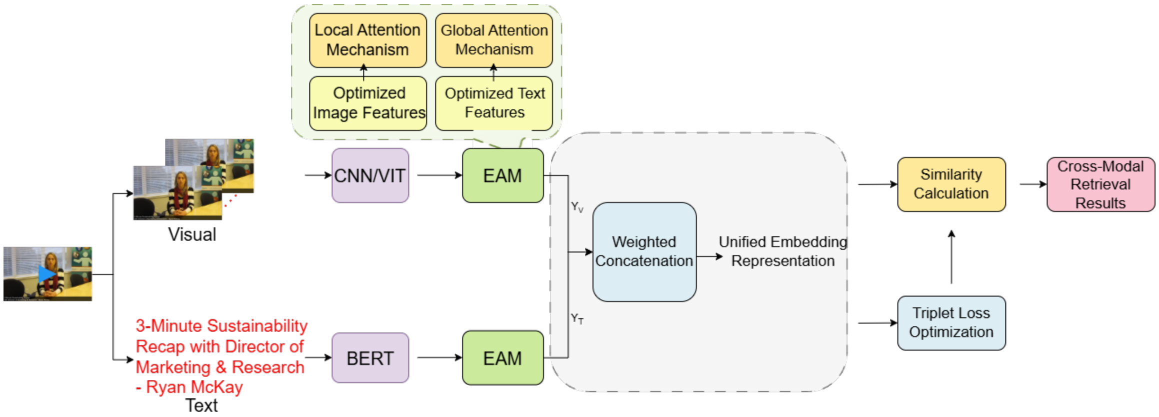
Local Attention Mechanism
Global Attention Mechanism
Optimized Image Features
Optimized Text Features
Visual
CNN/VIT
EAM
Y_V
Weighted Concatenation
Unified Embedding Representation
Similarity Calculation
Cross-Modal Retrieval Results
Triplet Loss Optimization
Y_T
3-Minute Sustainability Recap with Director of Marketing & Research - Ryan McKay
Text
BERT
EAM

Figure 1: The overall architecture of the MEFN framework

Figure 1 illustrates the detailed overall architecture of the MEFN framework, showing the data flow and interactions between modules.

3.2. Implementation Details

3.3. Feature Extraction

The feature extraction part consists of two components: image feature extraction and text feature extraction.

Images are first processed through CNNs (e.g., ResNet101) or ViT to extract multi-scale spatial features, outputting a feature tensor $F_{img} \in R^{H \times W \times C}$. Images are divided into regional blocks and encoded with positional information to retain both local details and global semantics.

Text is converted into context-aware word-level embeddings $T_{txt} \in R^d$ using BERT, which employs a multi-layer Transformer structure to model deep semantic relationships between words in sentences.

3.4. Entropy-Aware Modeling

To better guide feature alignment between images and text, we introduce the Modality Fusion Entropy (MFE) mechanism. This module dynamically assigns fusion weights to different features by calculating the mutual information between image and text features, generating optimized modal features, as

$$MFE = \sum_{i=1}^{N} \sum_{j=1}^{M} H(v_i, t_j) \tag{1}$$

in which v_i represents the i-th regional feature of the image, t_j is the j-th word vector in the text, and N, M denote the number of image blocks and text length, respectively. The entropy function $H(\cdot)$ measures the information uncertainty between them, optimizing the quality of feature representation.

To validate the effectiveness of the Modality Fusion Entropy (MFE) mechanism in image-text alignment, we constructed a mutual information entropy heatmap between image regions and text word vectors (Figure 2). The horizontal axis represents text words t_j, the vertical axis represents image regions v_i, and the color denotes the entropy $H(v_i, t_j)$, indicating the uncertainty of semantic matching. Lower entropy values suggest stronger semantic relevance and thus receive higher fusion weights; higher values indicate redundancy or weak correlation.

As shown, different image regions exhibit distinct responses to text words. For instance, v_1, v_4, v_6 show lower entropy values for certain tokens, suggesting strong semantic discriminability, while v_5 shows high entropy at t_7, possibly indicating background or irrelevant areas. The non-uniform distribution of entropy reflects the structural heterogeneity between image and text modalities and provides a basis for adaptive feature fusion.

Although this heatmap is generated from a single sample, its design is grounded in the mathematical definition of entropy and thus has strong theoretical generality. Consistent trends were observed across multiple samples: low-entropy regions contribute more to fusion, while high-entropy regions are suppressed. These findings confirm the MFE module's ability to reduce redundancy and enhance semantic relevance from an information-theoretic perspective, thereby improving the discriminability and robustness of the fused representation.

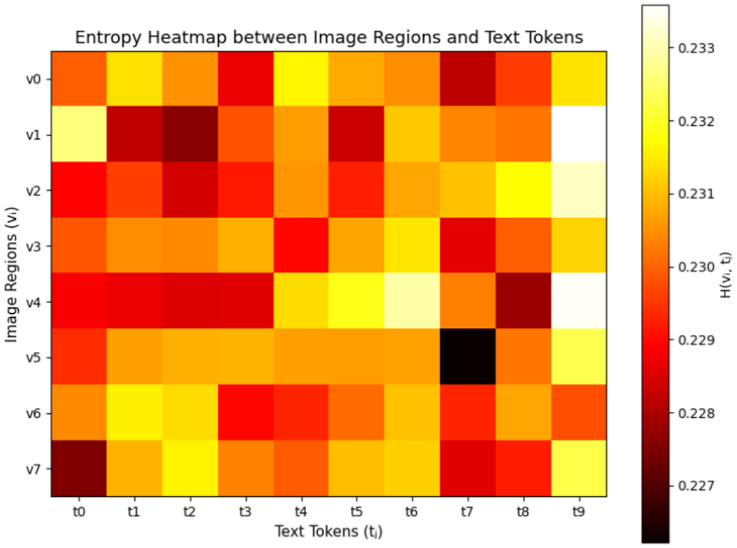

Figure 2: Mutual information entropy heatmap between image regions vi and text words tj.

3.5. Multi-Scale Attention Fusion

In this module, we apply local and global attention mechanisms to the optimized image and text features.

To more clearly illustrate the internal structure and information flow of the multi-scale attention fusion module, we present a class diagram in pseudocode form, as shown in Figure 3. The entire process consists of four key components: Input Features, Local Attention, Global Attention, and Weighted Fusion. The final output is a unified cross-modal representation f_fusion.

The Input Features module receives entropy-optimized image and text modality features F_img' and F_txt';

The Local Attention component computes the similarity sim(v_i, t_j) between image regions and text words, and generates the local semantic fusion output f_local based on attention weights alpha_ij;

The Global Attention component focuses on the semantic consistency between the global image representation v_global and the sentence-level text embedding t_cls, outputting the global semantic feature f_global;

The outputs of both branches are fed into the Weighted Fusion module, where they are fused via a function fuse_local, f_global to produce the final unified representation f_fusion, which serves as the cross-modal embedding vector for downstream retrieval tasks.

3.6. Image-Text Joint Representation and Retrieval

The unified embedding is input into the image-text matching space and trained using a triplet loss function, which maximizes the similarity between positive image-text pairs and minimizes it between negative pairs:

$$L_{triplet} = max(||f(v) - f(t)||_2^2 - ||f(v) - f(t^+)||_2^2 + \alpha, 0) \tag{2}$$

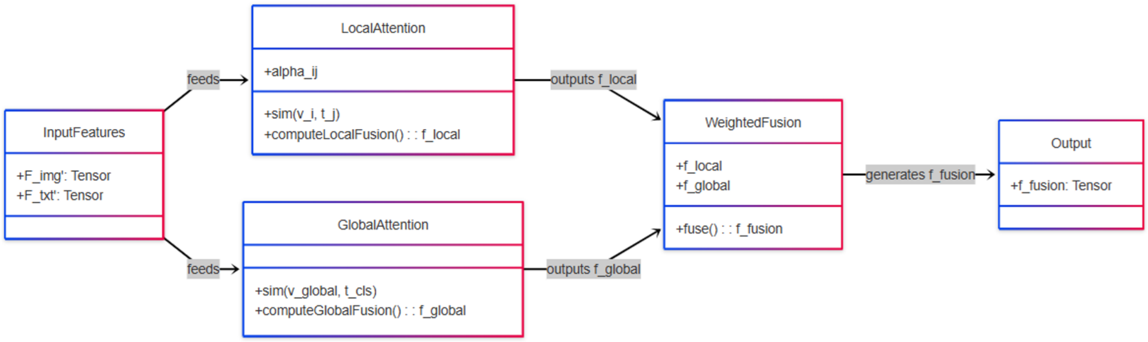

Figure 3: Class diagram representation of the multi-scale attention fusion module.

in which $f(\cdot)$ is the feature encoding function, $t+$ is a negative text sample, and α is the margin threshold. After training, the model efficiently calculates similarity and performs image-text matching to output cross-modal retrieval results.

4. Experimental Results

This part details the experimental design to validate the effectiveness of the MEFN method, focusing on datasets, evaluation metrics, comparison models, and the design and implementation of ablation experiments.

4.1. Datasets and Evaluation Metrics

To evaluate the performance of MEFN, we selected two widely used ITR datasets: Flickr30K and MSCOCO, which contain large numbers of images with corresponding natural language descriptions and serve as standard benchmarks for ITR models. Flickr30K includes 31,000 images, each with 5 text descriptions, used for image-text matching tasks. It is widely employed to test the capabilities of ITR models, especially for diverse image-description pairs. MSCOCO (Microsoft Common Objects in Context) is large-scale, containing 123,000 images with 5 natural language descriptions each. Its diversity and complexity make it a key benchmark in the ITR field.

The following common evaluation metrics are adopted. Recall@1 measures the proportion of correct matches in the top 1 retrieval result, assessing retrieval precision. Recall@5 is similar to Recall@1 but for the top 5 results, providing a more comprehensive evaluation of retrieval quality. Mean Average Precision (mAP) calculates the average precision across all retrieval results, reflecting the model's overall performance and widely used as a standard evaluation metric in ITR.

4.2. Baseline Models

To comprehensively evaluate MEFN, we selected the following ITR models as baselines.

A global alignment-based ITR method VSE++ matches images and text through global feature alignment, learning a visual-language space to improve correlation calculation. SCAN utilizes fine-grained alignment of image and text features, employing methods for fusing fine-grained information to significantly enhance performance in complex image-text matching tasks, especially for diverse descriptions.CLIP, a VLP-based ITR model advances cross-modal retrieval through joint training on large-scale datasets, pushing the boundaries of the field.

Comparisons with these models allow a comprehensive assessment of MEFN's advantages and improvements.

4.3. Experimental Result

To validate the effectiveness of MEFN, experiments were conducted on Flickr30K and MSCOCO using Recall@1, Recall@5, and mAP as evaluation metrics.

Table 1 presents the experimental results on Flickr30K. On Flickr30K, MEFN achieves 82.5% Recall@1, 94.3% Recall@5, and 76.2% mAP, surpassing VSE++ (78.1%, 92.7%, 72.8%) and SCAN (80.0%, 93.1%, 74.5%). This improvement indicates that MEFN better captures fine-grained semantic correlations between images and text, even when dealing with diverse image-description pairs. The entropy-aware modeling (MFE) likely mitigates noise from redundant background information in images, while the local attention mechanism aligns regional image features with specific text keywords enhancing precision in detailed matching.

Table 1: Experimental results on Flickr30K

model	Recall@1 (%)	Recall@5 (%)	mAP (%)
VSE++	78.1	92.7	72.8
SCAN	80.0	93.1	74.5
MEFN	**82.5**	**94.3**	**76.2**

Table 2 presents the experimental results on MSCOCO. On the more complex MSCOCO dataset, MEFN achieves 84.3% Recall@1, 95.5% Recall@5, and 78.9% mAP, significantly outperforming VSE++ (80.5%, 93.9%, 75.3%) and SCAN (81.2%, 94.2%, 76.7%). The superior performance in complex scenes (e.g., multi-object images with rich textual descriptions) highlights the effectiveness of the global attention mechanism in capturing overall semantic consistency. By integrating both local details and global context, MEFN resolves the multi-scale alignment challenge, ensuring that large-scale visual features (e.g., scene layouts) align with global text semantics, while fine-grained features (e.g., object attributes) match local text content.

Table 2: Experimental results on MSCOCO

model	Recall@1 (%)	Recall@5 (%)	mAP (%)
VSE++	80.5	93.9	75.3
SCAN	81.2	94.2	76.7
MEFN	**84.3**	**95.5**	**78.9**

4.4. Ablation Experiments

To analyze the contributions of each module in MEFN, the ablation experiments are conducted by removing key modules and evaluating their impact on model performance on MSCOCO.

Remove the Entropy-Aware Module (MFE) to evaluate the role of entropy-aware modeling in feature fusion. This module affects the model's feature fusion ability and impacting cross-modal alignment accuracy.Remove the Local Attention Mechanism to analyzes the contribution of local attention to fine-grained semantic matching, which aligns image details with text keywords. Remove the Global Attention Mechanism to investigate the impact of the global attention, which focuses on overall semantic alignment and may weaken global semantic correlation, affecting retrieval performance.

Results from each ablated experiment are compared with the full model to analyze the specific effects of each module, verifying their effectiveness and complementarity in the MEFN framework, shown in table 3. Removing MFE (MEFN-MFE) leads to a drastic drop in Recall@1 (71.2%) and mAP (65.3%), confirming that entropy-based feature weighting is essential for reducing noise in cross-modal fusion. Without MFE, the model struggles to distinguish informative features from redundant ones, impairing alignment accuracy.Removing local attention (MEFN-Local Attention) decreases Recall@1 to 75.8% and mAP to 69.8%, indicating that fine-grained alignment between image regions and text words is critical for detailed semantic matching (e.g., recognizing specific objects or actions). Removing global attention (MEFN-Global Attention) results in lower Recall@5 (87.1%) and mAP (70.2%), showing that global semantic consistency is necessary for capturing the overall context of images and text, especially in complex scenes with multiple semantic elements.

These results confirm that the synergy between entropy-aware modeling and dual-scale attention mechanisms is key to MEFN's performance, as each component addresses distinct challenges: MFE handles feature noise, local attention improves fine-grained alignment, and global attention ensures holistic semantic consistency.

Table 3: The results of on ablation experiments

model	Recall@1 (%)	Recall@5 (%)	mAP (%)
MEFN-MFE	71.2	83.5	65.3
MFEN-Local Attention	75.8	86.7	69.8
MFEN-Global Attention	76.3	87.1	70.2
MEFN	84.3	95.5	78.9

5. Conclusion

This paper addresses core challenges in cross-modal image-text retrieval, such as modality heterogeneity, information redundancy, and insufficient semantic alignment, by proposing MEFN (Multi-scale Entropy-aware Fusion Network). The method uses Modality Fusion Entropy (MFE) to guide information transfer between images and text, effectively mitigating noise interference in cross-modal feature coupling. Meanwhile, the dual-scale attention mechanism balances fine-grained local semantic alignment between image details and text keywords and global semantic consistency at the contextual level, enhancing the discriminative and expressive capabilities of cross-modal representations.

Experimental results on standard datasets like Flickr30K and MSCOCO demonstrate that MEFN outperforms mainstream methods in multiple metrics, achieving higher retrieval accuracy and stronger semantic robustness, especially maintaining stable performance under complex image scenarios and diverse text descriptions.

Acknowledgments

All the research in this paper is supported by Henan Provincial Key Research and Development Program in Science and Technology (232102210102), Key Research Project for Higher Education Institutions of Henan Province (25A520023), the Natural Science Foundation of Zhongyuan University of Technology (K2023QN019).

References

M. Cao, S. Li, J. Li, et al. Image-text retrieval: A survey on recent research and development. *arXiv preprint arXiv:2203.14713*, 2022. doi: 10.48550/arXiv.2203.14713.

F. L. Chen, D. Z. Zhang, M. L. Han, et al. Vlp: A survey on vision-language pre-training. *Machine Intelligence Research*, 20(1):38–56, 2023. doi: 10.1007/s11633-022-1369-5.

Z. Gan, L. Li, C. Li, et al. Vision-language pre-training: Basics, recent advances, and future trends. *Foundations and Trends® in Computer Graphics and Vision*, 14(3-4):163–352, 2022. doi: 10.48550/arXiv.2210.09263.

Z. Ji, K. Chen, Y. He, et al. Heterogeneous memory enhanced graph reasoning network for cross-modal retrieval. *Science China Information Sciences*, 65(7):172104, 2022. doi: 10.1007/s11432-021-3367-y.

P. P. Liang, Y. Cheng, R. Salakhutdinov, et al. Multimodal fusion interactions: A study of human and automatic quantification. In *Proceedings of the 25th International Conference on Multimodal Interaction*, pages 425–435, 2023. doi: 10.1145/3577190.3614151.

S. Mai, Y. Zeng, and H. Hu. Multimodal information bottleneck: Learning minimal sufficient unimodal and multimodal representations. *IEEE Transactions on Multimedia*, 25:4121–4134, 2022. doi: 10.1109/TMM.2022.3171679.

P. Shrestha, S. Amgain, B. Khanal, et al. Medical vision language pretraining: A survey. *arXiv preprint arXiv:2312.06224*, 2023. doi: 10.48550/arXiv.2312.06224.

J. M. Vala and U. K. Jaliya. Deep learning network and renyi-entropy based fusion model for emotion recognition using multimodal signals. *International Journal of Modern Education and Computer Science*, 11(4):67, 2022. doi: 10.5815/ijmecs.2022.04.06.

H. Wang, J. Du, Y. Dai, et al. Improving multi-modal emotion recognition using entropy-based fusion and pruning-based network architecture optimization. In *ICASSP 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11766–11770. IEEE, 2024. doi: 10.1109/ICASSP48485.2024.10447231.

T. Wang, X. Xu, Y. Yang, et al. Matching images and text with multi-modal tensor fusion and re-ranking. In *Proceedings of the 27th ACM international conference on multimedia*, pages 12–20, 2019. doi: 10.1145/3343031.3350875.

T. Wang, F. Li, L. Zhu, et al. Cross-modal retrieval: a systematic review of methods and future directions. *Proceedings of the IEEE*, 2025. doi: 10.1109/JPROC.2024.3525147.

S. Wu, D. Dai, Z. Qin, et al. Denoising bottleneck with mutual information maximization for video multimodal fusion. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 2231–2243, 2023. doi: 10.18653/v1/2023.acl-long.124.

R. Yang, S. Wang, Y. Han, et al. Transcending fusion: A multi-scale alignment method for remote sensing image-text retrieval. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. doi: 10.1109/TGRS.2024.3496898.

C. Zhang, J. Song, X. Zhu, et al. Hcmsl: Hybrid cross-modal similarity learning for cross-modal retrieval. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 17(1s):1–22, 2021. doi: 10.1145/3412847.

Q. Zhang, Z. Lei, Z. Zhang, et al. Context-aware attention network for image-text retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3536–3545, 2020. doi: 10.1109/CVPR42600.2020.00359.

K. Zhou, F. H. Hassan, and G. K. Hoon. The state of the art for cross-modal retrieval: A survey. *IEEE Access*, 11:138568–138589, 2023. doi: 10.1109/ACCESS.2023.3338548.

H. Zhu, Y. Wei, Y. Zhao, et al. Amc: Adaptive multi-expert collaborative network for text-guided image retrieval. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(6):1–22, 2023. doi: 10.1145/3584703.