Composed Image Retrieval Based on Subject Feature Extraction

Yang Ning

***Abstract*— Composed image retrieval, a critical computer vision task enhancing retrieval via multi-modal fusion, faces challenges in accurately extracting subject features and minimizing irrelevant interference during fusion; this paper addresses these by proposing a subject feature extraction-based method integrating text-guided segmentation and multi-modal fusion. The approach combines CLIP's cross-modal alignment with Swin Transformer's hierarchical feature learning, dynamically focusing segmentation on text-relevant regions via cross-modal attention and skip-layer fusion to generate precise masks for visual features. A two-stage framework first filters irrelevant image details through segmentation, then uses bidirectional multi-head cross-attention in an image-text fusion module to enable fine-grained interactions, decouple redundant semantics, and reinforce discriminative feature correlations. Validated on FashionIQ and CIRR datasets, the method demonstrates improved retrieval accuracy, with segmentation preserving text-relevant details and fusion enhancing semantic consistency, offering robust solutions for e-commerce and security while advancing multi-modal feature decoupling and alignment.**

***Index Terms*—compositional image retrieval, subject feature extraction, text-guided segmentation, multi-modal fusion.**

I. Introduction

Composed Image Retrieval (CIR) aims to achieve precise retrieval through the joint constraints of reference images and textual descriptions. Its core task involves matching target images via text-guided descriptions of reference images. Due to the need to align cross-modal semantic differences between reference and target images, researchers have recently proposed various technical frameworks to overcome the limitations of traditional unimodal retrieval. Current mainstream methods perform retrieval by simply fusing global image features with textual description features. However, these approaches struggle to accurately extract semantically relevant image subject features during image processing, leading to retrieval results often deviating from users' actual requirements. Additionally, during feature fusion, they fail to effectively distinguish between text-relevant and irrelevant information in images, resulting in fusion features contaminated by substantial irrelevant data and hindering improvements in retrieval precision. Therefore, this paper focuses on extracting subject features and managing irrelevant information.

To address the challenge of accurately extracting semantically aligned image subject features in CIR, a text-guided image segmentation method is proposed to isolate image subjects for subsequent feature extraction. This method uses the CLIP model to encode user descriptions into semantic embedding vectors, then employs attention mechanisms to interact text semantic embeddings with image features, dynamically guiding the segmentation network to focus on text-relevant regions and generating segmentation masks consistent with textual semantics. This provides precise visual information for cross-modal feature fusion. To tackle irrelevant information in fused features degrading retrieval precision, a compositional retrieval method based on subject feature extraction is introduced. It leverages text-guided segmentation to isolate text-relevant regions and suppress irrelevant backgrounds, then fuses segmented image features with text features to generate discriminative retrieval vectors. These vectors are further combined with text features to enhance retrieval accuracy and efficiency.

Manuscript received March 21, 2025

Yang Ning, School of computer science and technology, Tiangong University, Tianjin, China

II. Related Works

A. ResNet

ResNet[1] (Residual Network), proposed by Kaiming He et al. in 2015, is a revolutionary convolutional neural network (CNN) architecture. Its core innovation is the introduction of residual connections to address the gradient vanishing and degradation problems in deep neural networks. By allowing information to bypass layers through "shortcut connections," ResNet enables the training of extremely deep models (e.g., over 100 layers) without performance degradation.

A residual block in ResNet has the form ($F(x) = x + H(x)$), where x is the input, and ($H(x)$) is the learned residual function. This design preserves original features while learning incremental improvements, significantly enhancing training stability and accuracy.ResNet achieved groundbreaking results in the ImageNet competition, demonstrating that deeper networks can surpass prior state-of-the-art models when properly optimized. It has since become a foundational backbone for various computer vision tasks, including image classification, object detection, and semantic segmentation.

B. Swin Transformer

Developed by Microsoft Research Asia in 2021, Swin Transformer[2] is a vision Transformer model that addresses computational bottlenecks in traditional Transformers through its hierarchical architecture and shifted window mechanism. Mimicking CNN's pyramidal structure, its hierarchical design generates multi-scale features via progressive downsampling, enabling the model to capture both local details and global context, which inherently suits multi-task scenarios like object detection and segmentation. The shifted window mechanism restricts self-attention calculations to local windows, avoiding quadratic complexity

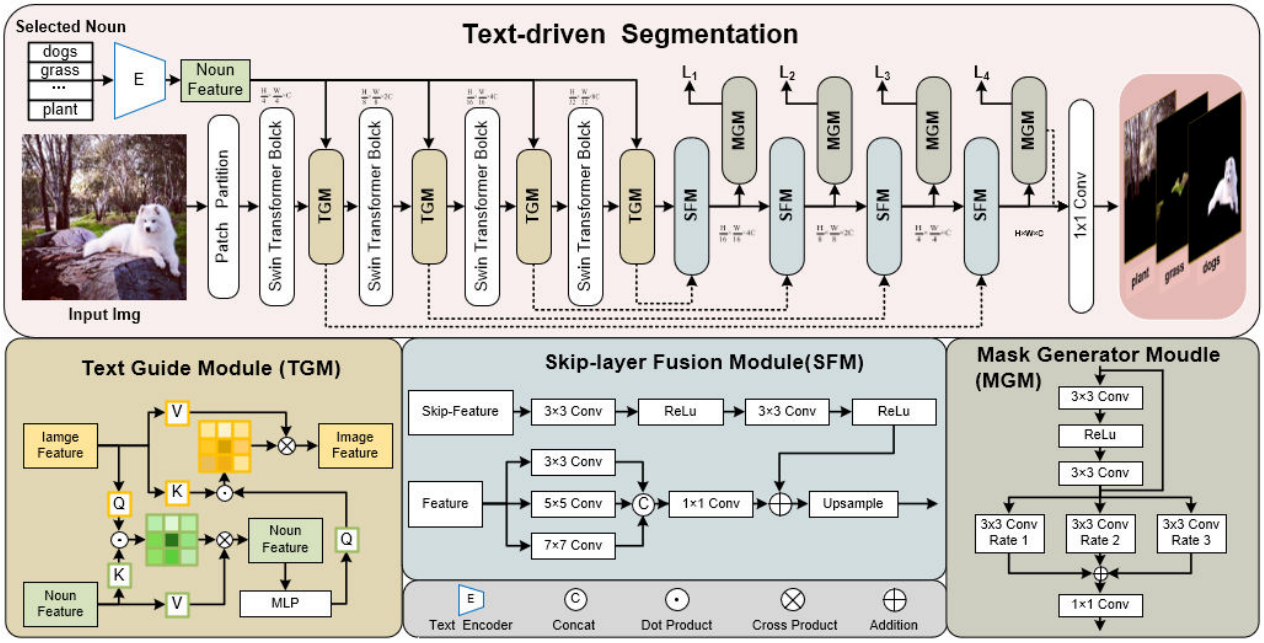

Fig 1 Text-driven Segmentation Framework

and ensuring efficiency in processing high-resolution images, while window shifting allows information exchange between adjacent regions to balance local and global modeling. Additionally, its modular architecture supports flexible adjustments to window size, depth, and channel dimensions, optimizing model scalability for diverse tasks and hardware constraints. These attributes establish Swin Transformer as a versatile backbone for vision tasks, driving the widespread adoption of Transformers in real-world applications.

C. CLIP

Developed by OpenAI in 2021, CLIP[3] is a groundbreaking multi-modal model that bridges language and vision through contrastive learning, embedding images and text into a unified semantic space for cross-modal understanding and generation. Its core design leverages large-scale unsupervised pretraining on publicly available internet image-text pairs, learning generalizable feature representations by maximizing the similarity of matched image-text pairs and minimizing that of mismatched ones. This training strategy endows CLIP with robust zero-shot transfer capabilities, enabling it to perform tasks like image classification and object detection without task-specific fine-tuning, even for unseen concepts. Compared to traditional models, CLIP significantly reduces reliance on labeled data, driving advancements in open-domain multi-modal applications such as cross-lingual retrieval and image generation. Its architecture and training philosophy have laid the foundation for subsequent multi-modal large-scale models.

D. DeepLabv3+

DeepLabv3+[4], proposed by Google in 2018, is a semantic segmentation model that enhances the original DeepLabv3 by introducing an encoder-decoder architecture to balance global semantic context and local detail extraction. Its key innovations include: an encoder leveraging depthwise separable convolutions in backbone networks (e.g., Xception[5]) alongside multi-scale dilated convolutions via the Atrous Spatial Pyramid Pooling (ASPP) module to capture objects of varying sizes; a decoder that restores spatial resolution through upsampling and fuses low-level high-resolution features to improve boundary localization; and a lightweight design minimizing computational redundancy for real-time inference. Demonstrating superior performance on benchmarks like PASCAL VOC, DeepLabv3+ is widely applied in autonomous driving, medical image analysis, and other scenarios requiring precise scene understanding.

III. Methods

A. Text-driven Segmentation

Our approach combines multi-modal pretrained models CLIP and Swin Transformer, consisting of three integrated modules: the Text Guide Module (TGM) uses cross-attention to embed text semantics into image features, enabling text-driven guidance; the Skip-layer Fusion Module (SFM) decodes fused features by leveraging skip connections and multi-receptive-field aggregation to enhance spatial detail retention; and the Mask Generator Module (MGM) employs dilated convolutions to capture multi-scale image context, generating hierarchical masks for loss computation and training optimization. This architecture synergistically aligns textual guidance with visual feature learning, ensuring precise semantic localization while maintaining computational efficiency.

1) Text Guide Module (TGM)

The TGM dynamically aligns textual and visual features through CLIP and Swin Transformer collaboration, consisting of two phases. First, the CLIP text encoder extracts text features F_t, while input images are downsampled via Patch Partition and processed through four Swin Transformer stages to generate hierarchical image features F_1-F_4 with resolutions $\frac{H}{4}\times\frac{W}{4},\frac{H}{8}\times\frac{W}{8},\frac{H}{16}\times\frac{W}{16},\frac{H}{32}\times\frac{W}{32}$. In the text guidance phase, bidirectional cross-attention mechanisms are applied at each Swin Transformer stage: image features Q_i

interact with text K_t/V_t as:

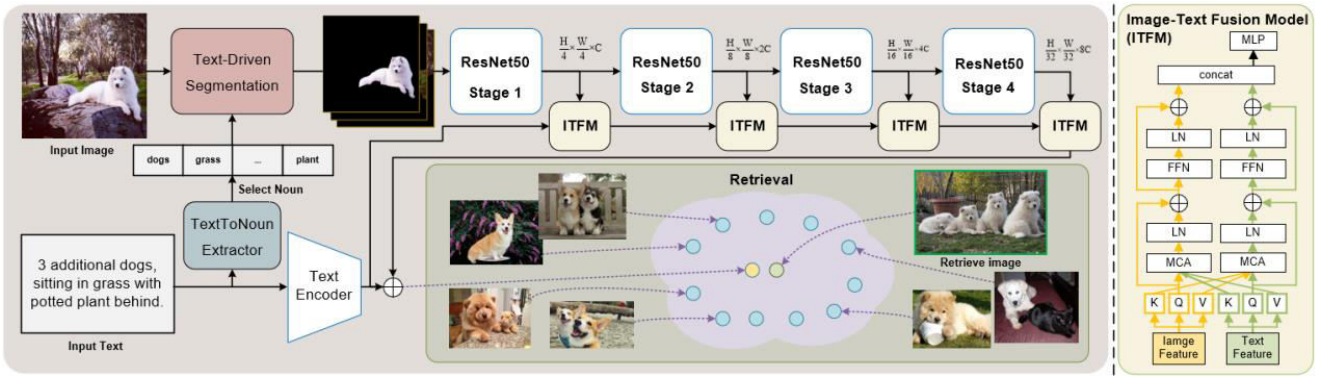

Fig 2 Composed Image Retrieval Framework

$$Attention(Q_i, K_t, V_t) = Softmax\left(\frac{Q_i K_t^T}{\sqrt{d_k}}\right)V_t \quad (1)$$

generating text-guided features mapped to Q_t through an MLP. These Q_t then recompute attention with image K_i/V_i as:

$$Attention(Q_t, K_i, V_i) = Softmax\left(\frac{Q_t K_i^T}{\sqrt{d_k}}\right)V_i \quad (2)$$

producing refined text-guided image features $F_1' - F_4'$ provide semantically aligned visual inputs for downstream segmentation tasks, improving precision through progressive text-image alignment.

2) Skip-layer Fusion Module (SFM)

The SFM integrates multi-scale text-guided features $F_1' - F_4'$ via parallel convolutions (3×3, 5×5, 7×7), channel concat -enation, and skip connections. It captures local context through hierarchical receptive fields, reduces dimensions with 1×1 convolutions, and preserves details via residual connections. Upsampling restores spatial resolution for decoder-stage alignment, iteratively refining features $F_1'' - F_4''$. This design enhances discriminative multi-scale represent-tations, balancing global semantics and local details to improve segmentation accuracy.

3) Mask Generator Module (MGM)

The MGM generates semantically aligned segmentation masks using multi-scale dilated convolutions and multi-level supervision. Taking hierarchical features($F_1'' - F_4''$)from the skip-layer fusion module, each feature map undergoes 3×3 convolution, followed by parallel dilated convolutions (dilation rates 1, 2, 3) to capture multi-scale context. Features are then dimension-reduced via 1×1 convolution to produce corresponding masks ($Mask_1 - Mask_4$) at their respective resolutions. During training, all masks contribute to loss calculation, enforcing fine-grained text-image alignment across scales. At inference, only the highest-resolution mask ($Mask_1$) is used for segmentation, while others ($Mask_2 - Mask_4$) provide auxiliary supervision. This approach enhances boundary precision and semantic consistency through multi-scale aggregation and hierarchical supervision, ensuring reliable region localization for compositional retrieval tasks.

4) Loss Function

Inspired by DeepLabV3+, our segmentation loss combines Cross Entropy (CE) and Dice Loss through a joint optimization strategy to enhance robustness in text-guided segmentation. The CE loss：

$$L_{CE} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log(p_{i,c}) \quad (3)$$

minimizes pixel-wise prediction errors for fine-grained classification, while Dice Loss :

$$L_{Dice} = 1 - \frac{2\sum_{i=1}^{N} y_i p_i}{\sum_{i=1}^{N} y_i + \sum_{i=1}^{N} p_i + \epsilon} \quad (4)$$

optimizes region overlap for semantic consistency. The total loss is defined as:

$$L_{total} = \alpha L_{CE} + \beta L_{Dice} \quad (5)$$

with experimentally determined weights α = 1.0 and β = 0.5. CE prioritizes pixel-level classification accuracy, while Dice Loss reinforces boundary integrity. This combination improves semantic alignment and spatial precision of segm -entation masks, balancing local detail and global structure optimization.

B. Composed Image Retrieval

The model is composed of three processes: keyword extraction using Spacy, the multi-modal pre-trained CLIP, and text-guided image segmentation. The compositional image retrieval process includes text keyword extraction, text-guided segmentation, image-text feature fusion, and retrieval feature generation. First, the query image and text are input. Spacy extracts nouns from the text as keywords to guide image segmentation, generating a semantically consistent segmented image and reducing interference from irrelevant backgrounds. The segmented image is then fed into the CLIP image encoder to extract features. The output features of its four levels are fused with the text features extracted by CLIP via the ITFM module. The ITFM module uses an attention mechanism and a multi-layer perceptron (MLP) to adjust feature dimensions, obtaining fused features at four levels. In the retrieval feature generation stage, the final retrieval features are generated using the formula：

$$F_{retrieval} = F_{text} + \sum_{i=1}^{4} F_{fusion}^{i} \quad (6)$$

Where F_{text} is the original text feature encoded by CLIP and F_{fusion}^{i} is the fused feature at the i -th level. During retrieval, the similarity between the retrieval features and the image features in the retrieval library is calculated and sorted to return results, improving retrieval accuracy. The ITFM module achieves fine-grained alignment of image and text features. First, query, key, and value vectors are generated from the text feature F_{text} and the image feature F_{image} through linear transformation. The multi-head cross-attention

mechanism (MCA) is adopted. The cross-attention of image features to text features is calculated as:

$$CrossAttention(q_i, k_t, v_t) = Softmax\left(\frac{q_i k_t^T}{\sqrt{d_k}}\right) v_t \quad (7)$$

and that of text features to image features is:

$$CrossAttention(q_t, k_i, v_i) = Softmax\left(\frac{q_t k_i^T}{\sqrt{d_k}}\right) v_i \quad (8)$$

The MCA output is concatenated and linearly transformed as:

$$MCA(q_i, k_t, v_t) = W_0 Concat(head_1, head_2, \dots, head_h) \quad (9)$$

and then goes through layer normalization (LN) and a feed-forward network (FFN) to get the intermediate features:

$$F_i = FFN\left(LN\left(MCA(q_i, k_t, v_t)\right)\right) \quad (10)$$

$$F_t = FFN\left(LN\left(MCA(q_t, k_i, v_i)\right)\right) \quad (11)$$

Finally, F_i and F_t are concatenated in the channel dimension and the dimension is adjusted by MLP:

$$F_{fusion} = MLP\left(Concat(F_i, F_t)\right) \quad (12)$$

This fused feature supports subsequent retrieval and enhances the discriminative ability of retrieval features.

C. Loss Function

To optimize retrieval tasks, this paper employs BBCL, which outperforms Triplet Loss on complex datasets with superior discriminative power and faster convergence. Each training batch contains B query pairs, each consisting of a reference image, modified text, and their target image. The loss function is defined as:

$$L_{BBCL} = \frac{1}{B}\sum_{i=1}^{B} -\log\left(\frac{\exp\left(\kappa(q_i, t_i)\right)}{\sum_{j=1}^{B}\exp\left(\kappa(q_i, t_j)\right)}\right) \quad (13)$$

Here, q_i denotes the fused features of the i-th query sample, t_i represents the features of the i-th target image, and $\kappa(\cdot,\cdot)$ is an arbitrary distance metric function (e.g., cosine distance). Combining the segmentation loss from Chapter 3, the overall loss function of this method is:

$$L_{retrieval} = L_{BBCL} + L_{total} \quad (14)$$

IV. Experiment

A. Dataset

The experiments utilize two datasets: FashionIQ [50] and CIRR [24]. FashionIQ, a benchmark for natural language-based image retrieval, contains 77,684 fashion product images (Dress, Toptee, Shirt) organized into triplets: reference image, relative text description, and target image. These triplets illustrate attribute modifications between the reference and target images. CIRR addresses limitations in domain-specific datasets like FashionIQ by introducing 21,552 real-world images from the NLVR² dataset, forming 36,554 triplets with 80%-10%-10% splits. Its design mitigates visual complexity constraints and reduces false negatives by leveraging real-life scenarios. Both datasets are critical for evaluating the model's compositional retrieval performance across controlled fashion and diverse real-world contexts.

B. Evaluation Metrics

Recall, also known as the true positive rate, is calculated as:

$$Recall = \frac{TP}{TP + FN} \quad (15)$$

Here, TP (True Positive) represents the number of positive instances correctly identified by the model, while FN (False Negative) refers to positive instances misclassified as negative. As shown in the formula, Recall measures the proportion of actual positive samples that the model successfully retrieves.

C. Experimental Results

This section evaluates the performance of our method on the FashionIQ[15] and CIRR[9] datasets through quantitative experiments, comparing it with existing approaches. For FashionIQ, we use Recall@K as the evaluation metric, focusing on Recall@10 (R@10) and Recall@50 (R@50). On CIRR, following the standard evaluation protocol proposed by the dataset authors, we report Recall@K at four different ranks (1, 5, 10, 50). Additionally, leveraging CIRR's unique design, we report RecallSubset, which considers only images within the query subset. Since the ground-truth labels of the test set were not publicly available during manuscript preparation, all experimental results were computed on the validation set. The results demonstrate that our method achieves significant performance improvements across multiple evaluation metrics.

1) CIRR

On the CIRR dataset (Table 1), our method achieves 44.21%/ 76.57%/ 86.67%/ 97.88% in Recall@1/5/10/50, representing +5.24%/ +3.95%/ +4.70%/ +1.15% gains over the best baseline SSN. Subset evaluation metrics (R_subset@1/2/3: 71.37%/ 88.14%/ 94.91%) show +5.41%/ +3.34%/ +1.52% improvements compared to SSN, demonstrating the model's robust fine-grained retrieval capability in scenarios with high-similarity distractors.

Table 1 CIRR result

	Recall@K				$Recall_{Subset}$@K		
Methods	K=1	K=5	K=10	K=50	K=1	K=2	K=3
TIGR[6]	14.61	48.37	64.08	90.03	22.67	44.97	65.14
MAAF[7]	10.31	33.03	48.30	80.06	21.05	41.81	61.60
ARTEMIS[8]	16.96	46.10	61.31	87.73	39.99	62.20	75.67
CIRPLANT[9]	19.55	52.55	68.39	92.38	39.20	63.03	79.49
CLIP4Cir[10]	33.59	65.35	77.35	95.21	62.39	81.81	92.02
SSN[11]	38.97	72.62	81.97	96.73	65.96	84.80	93.39
Our	**44.21**	**76.57**	**86.67**	**97.88**	**71.37**	**88.14**	**94.91**

2) FashionIQ

On the FashionIQ dataset (Table 2), our method achieves 42.91% average Recall@10 across Dress/ Toptee/ Shirt categories (with the highest single-category value of 46.65%), representing a 6.41% improvement over SSN.

For Recall@50, the average reaches 68.69% (highest single-category: 72.13%), outperforming SSN by 6.04%. These results highlight the model's ability to capture fine-grained text-image semantic alignment, significantly enhancing retrieval accuracy.

Table 2 FashionIQ result

	Dress		Toptee		Shirt		Average	
Methods	R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50
CoSMo[12]	25.64	50.30	29.21	57.46	24.90	49.18	26.58	52.32
DCNet[13]	28.95	56.07	30.44	58.29	23.95	47.30	27.78	53.89
CIRPLANT	17.45	40.41	21.64	45.38	17.53	38.81	18.87	41.53
SAC[14]	26.52	51.01	32.70	61.23	28.02	51.86	29.08	54.70
MAAF	23.80	48.60	27.90	53.60	21.30	44.20	24.30	48.80
CLIP4Cir	31.63	56.67	38.19	62.42	36.36	58.00	35.39	59.03
SSN	32.18	58.05	41.01	66.12	36.33	63.80	36.50	62.65
Our	**38.52**	**63.16**	**43.57**	**72.13**	**46.65**	**70.78**	**42.91**	**68.69**

3) Ablation

Ablation studies on the CIRR dataset (Table 3) revealed that the Text Guide and Image-Text Fusion Module (IT -FM) had the most significant impact on performance. The full model, leveraging the synergy of text-guided segmen -tation, skip-layer fusion, mask generation, and cross-modal feature fusion, outperformed all variant models.

Table 3 Ablation studies on the CIRR

	Recall@K				RecallSubset@K		
Moudle	K=1	K=5	K=10	K=50	K=1	K=2	K=3
baseline	33.59	65.35	77.35	95.21	62.39	81.81	92.02
w/o Text Guide	41.78	76.23	86.37	97.63	68.38	85.64	94.04
w/o Skip-layer	42.43	76.29	86.24	97.77	69.21	86.41	94.21
w/o Mask Generator	42.64	76.63	86.10	97.65	69.07	86.56	94.28
w/o Text-Driven	40.23	74.81	85.19	97.27	66.85	84.36	93.75
w/o Image-Text	40.90	75.46	86.39	97.20	67.16	85.52	93.87
Full Model	44.21	76.57	86.67	97.88	71.37	88.14	94.91

V. Conclusion

Composed image retrieval faces challenges in accurately modeling user intent due to difficulties in extracting semantically aligned subject features and severe interference from irrelevant features during fusion. This paper addresses these issues by proposing a subject feature extraction-based method that combines text-guided segmentation and multi-modal feature fusion to enhance retrieval accuracy and efficiency. The approach first leverages the CLIP model to extract text semantic embeddings, then uses a cross-modal attention module to dynamically guide a segmentation network to focus on text-related regions. By integrating hierarchical features from Swin Transformer with a skip-layer fusion mechanism, pixel-level semantic alignment is achieved, providing high-purity visual features for subsequent retrieval tasks. Additionally, a text-guided segmentation module filters irrelevant background information while an Image-Text Fusion Module (ITFM) employs bidirectional multi-head cross-attention to enable fine-grained interactions between visual and textual features, generating discriminative retrieval features that reduce interference from non-subject elements. This synergy between Swin Transformer's hierarchical architecture, cross-modal attention, and text-guided segmentation bridges the gap between text and image semantics, advancing compositional retrieval in complex scenarios.

References

[1] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.
[2] Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 10012-10022.
[3] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]//International conference on machine learning. PmLR, 2021: 8748-8763.
[4] Chen L C, Zhu Y, Papandreou G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 801-818.
[5] Chollet F. Xception: Deep learning with depthwise separable convolutions[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 1251-1258.
[6] Vo N, Jiang L, Sun C, et al. Composing text and image for image retrieval-an empirical odyssey[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 6439-6448.
[7] Dodds E, Culpepper J, Herdade S, et al. Modality-agnostic attention fusion for visual search with text feedback[J]. arXiv preprint arXiv:2007.00145, 2020.
[8] Delmas G, de Rezende R S, Csurka G, et al. Artemis: Attention-based retrieval with text-explicit matching and implicit similarity[J]. arXiv preprint arXiv:2203.08101, 2022.
[9] Liu Z, Rodriguez-Opazo C, Teney D, et al. Image retrieval on real-life images with pre-trained vision-and-language models[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 2125-2134.
[10] Baldrati A, Bertini M, Uricchio T, et al. Conditioned and composed image retrieval combining and partially fine-tuning clip-based features[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 4959-4968.
[11] Yang X, Liu D, Zhang H, et al. Decomposing semantic shifts for composed image retrieval[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2024, 38(7): 6576-6584.
[12] Lee S, Kim D, Han B. Cosmo: Content-style modulation for image retrieval with text feedback[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 802-812.

[13] Kim J, Yu Y, Kim H, et al. Dual compositional learning in interactive image retrieval[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2021, 35(2): 1771-1779.
[14] Jandial S, Badjatiya P, Chawla P, et al. SAC: Semantic attention composition for text-conditioned image retrieval[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2022: 4021-4030.
[15] Wu H, Gao Y, Guo X, et al. Fashion iq: A new dataset towards retrieving images by natural language feedback[C]//Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition. 2021: 11307-11317.