DOI: 10.19884/j.1672-5220.202405006

Region-Aware Fashion Contrastive Learning for Unified Attribute Recognition and Composed Retrieval

WANG Kangping, ZHAO Mingbo*
College of Information Science and Technology, Donghua University, Shanghai 201620, China

Abstract: Clothing attribute recognition has become an essential technology, which enables users to automatically identify the characteristics of clothes and search for clothing images with similar attributes. However, existing methods cannot recognize newly added attributes and may fail to capture region-level visual features. To address the aforementioned issues, a region-aware fashion contrastive language-image pre-training (RaF-CLIP) model was proposed. This model aligned cropped and segmented images with category and multiple fine-grained attribute texts, achieving the matching of fashion region and corresponding texts through contrastive learning. Clothing retrieval found suitable clothing based on the user-specified clothing categories and attributes, and to further improve the accuracy of retrieval, an attribute-guided composed network (AGCN) as an additional component on RaF-CLIP was introduced, specifically designed for composed image retrieval. This task aimed to modify the reference image based on textual expressions to retrieve the expected target. By adopting a transformer-based bidirectional attention and gating mechanism, it realized the fusion and selection of image features and attribute text features. Experimental results show that the proposed model achieves a mean precision of 0.663 3 for attribute recognition tasks and a recall@10 (recall@*k* is defined as the percentage of correct samples appearing in the top *k* retrieval results) of 39.18 for composed image retrieval task, satisfying user needs for freely searching for clothing through images and texts.

Key words: attribute recognition; image retrieval; contrastive language-image pre-training (CLIP); image text matching; transformer

CLC number: TP391.4 **Document code**: A

Article ID: 1672-5220(2024)04-0405-11

Open Science Identity (OSID)

0 Introduction

Fashion is a multi-billion dollar industry with significant social, cultural, and economic effects. In recent years, the large number of clothing images on e-commerce websites and mobile applications has made it difficult for customers to find the right clothing items due to the wide range of colors, textures, and styles. To address this issue, clothing attribute recognition and retrieval technologies are applied to help users automatically identify clothing attributes and search for clothing that meets their needs.

Currently, most existing clothing attribute recognition methods[1-6] have adopted a deep learning pipeline for clothing image attribute prediction. However, a key problem with these methods is that they cannot handle arbitrary attribute predictions, especially in the presence of a new attribute. To address this problem, Radford et al.[7] proposed contrastive language-image pre-training (CLIP) to establish a connection between images and texts. The core idea was to map images and texts into the same embedding space and train the model by contrasting the similarities and differences between different images and texts, thus learning feature representations with good generalization ability. This method could also be applied to the field of clothing image-text retrieval. For example, Chia et al.[8] proposed FashionCLIP which fine-tuned the CLIP model using clothing images and their corresponding textual descriptions for zero-shot category recognition and multi-modal retrieval. However, there may be more than one piece of clothing in one image. In these cases, when matching with the text of a certain piece of clothing, ambiguity arises from not capturing the fine-grained alignment between the certain clothing regions and text spans.

In practical applications, retrieval systems often fail to retrieve suitable items in one search based on user needs, as it is difficult to express search intentions through a single image or text. In the field of fashion, composed image retrieval modifies the clothing attributes in the reference image according to textual descriptions, such as the dress length or the color of the clothing, to match the target image. In this task, the training data consists of multiple triplets, i.e., reference image, text and target image. Text image residual gating (TIRG) is the first work to fuse the reference image and textual modification using a residual gating mechanism[9].

Received date: 2024-05-15
Foundation item: National Natural Science Foundation of China (No. 61971121)
* Correspondence should be addressed to ZHAO Mingbo, email: mzhao4@dhu.edu.cn
Citation: WANG K P, ZHAO M B. Region-aware fashion contrastive learning for unified attribute recognition and composed retrieval[J]. *Journal of Donghua University (English Edition)*, 2024, 41(4): 405-415.

However, this method lacks interaction between visual information and textual information. To address this issue, visiolinguistic attention learning (VAL) inserts a synthesizer into the visual branch to preserve and transform visual representations based on textual modifications[10]. Content-style modulation (CoSMo) uses styles and contents to model the triplet relationship between images and texts[11]. Dual composition network (DCNet) considers the two circular mappings of the triplet by using a synthesis network and a correction network[12]. Semantic attention composition (SAC) employs a semantic feature attention and semantic feature modification module[13]. FashionViL further adopts a visual and linguistic representation framework in the fashion domain for multi-view contrastive learning and attribute classification[14]. Comprehensive linguistic-visual composition network (CLVC-Net) uses two cross-multi-granularity composition modules for multi-modal syntheses[15]. Cross relation network (CRN) models text features using a hierarchical aggregation transformer with a cross-relation network[16]. Image retrieval with text manipulation by local feature modification (LFM-IR) utilizes a spatial attention module and a channel attention module to facilitate semantic mapping between images and texts[17]. Feature extraction networks based on masked learning use masked auto-encoders and vision transformers to extract image descriptors and they are fine-tuned using self-supervised at the same time[18]. However, these methods semantically modify the reference image based on textual meaning, failing to consider that image features that have not been modified by the text still need to be preserved, which is unable to correctly use textual information to selectively modify the reference image features.

To address the aforementioned issues, this research proposes a region-aware fashion contrastive language-image pre-training (RaF-CLIP) model and an attribute-guided composed network (AGCN). Research has shown that removing the background can assist in fine-grained classification tasks, as it effectively eliminates background noise that affects prediction precision[19]. Based on this idea, this research first introduces the RaF-CLIP model, which aligns cropped and segmented region-level clothing images with categories and multiple fine-grained attribute texts, making the model suitable for clothing attribute recognition task. To achieve the composed image retrieval task, the AGCN network is designed based on RaF-CLIP. This network utilizes bidirectional attention and gating mechanisms to achieve selective modification of image representations by text, allowing for the iterative finding of results that meet user needs. The proposed model opens up new avenues for clothing retrieval in practical applications, not only improving retrieval accuracy but also enhancing user experience. It would have broad application prospects in e-commerce, social media and other related fields.

1 Model Disign

1.1 RaF-CLIP

Contrastive learning is a powerful self-supervised machine learning technique aimed at constructing an embedding space where semantic concepts from different modalities are similar. The objective of this research is to fine-tune the CLIP model while identifying both the category and fine-grained attributes of clothing. The proposed RaF-CLIP is illustrated in Fig. 1. For category recognition, the cropped image $\boldsymbol{I}$ based on the bounding box is fed into the image encoder V_θ to obtain visual representations $\boldsymbol{v} = V_\theta(\boldsymbol{I})$. The correct text is "a photo of (color) (category)", where "color" is extracted from the color recognition module that identifies the dominant color of the image and matches it to one of nine colors (red, orange, yellow, green, cyan, blue, purple, black and gray), providing a color cue for the text input. The corresponding text input $\boldsymbol{T}$ is tokenized and fed into the text encoder G_φ, resulting in textual representations $\boldsymbol{t} = G_\varphi(\boldsymbol{T})$. For each image-text feature pair $\beta = \{(\boldsymbol{v}_i, \boldsymbol{t}_i)\}_{i=1}^{L}$, where L is the size. The optimization objective is to maximize the cosine similarity between $\boldsymbol{v}_i$ and $\boldsymbol{t}_i$ while minimizing the cosine similarity between $\boldsymbol{v}_i$ and $\boldsymbol{t}_j$, $\forall\, i \neq j$, and i and j represent the number of image-text pairs. The Info-NCE[20] loss is used to supervise this task and the loss function L_c is specifically defined as:

$$L_c = \frac{1}{2}(L_{I2T} + L_{T2I}), \tag{1}$$

$$L_{I2T} = -\frac{1}{L}\sum_{i=1}^{L}\log\frac{\exp(\tau \boldsymbol{v}_i^{\mathrm{T}}\boldsymbol{t}_i)}{\sum_{j=1}^{L}\exp(\tau \boldsymbol{v}_i^{\mathrm{T}}\boldsymbol{t}_j)}, \tag{2}$$

$$L_{T2I} = -\frac{1}{L}\sum_{i=1}^{L}\log\frac{\exp(\tau \boldsymbol{t}_i^{\mathrm{T}}\boldsymbol{v}_i)}{\sum_{j=1}^{L}\exp(\tau \boldsymbol{t}_i^{\mathrm{T}}\boldsymbol{v}_j)}, \tag{3}$$

where L_{I2T} is the image-to-text (I2T) loss; L_{T2I} is the text-to-image (T2I) loss; τ is the temperature parameter.

In addition, a similar branch is utilized to train the model for identifying clothing attributes. The text input is a set of phrases "a photo of (attribute) (category)." The "attributes" include properties such as textures, fabrics and sleeve lengths of the clothing. The image input is fed into a segmentation network segment anything model (SAM)[21], utilizing the ground truth bounding boxes as prompts, to obtain images that only contain the clothing. The contrastive learning model trained with masked images focuses on the clothing regions, avoiding other noises, thus enabling the model to achieve better attribute recognition performance. After calculating the losses for the same image feature with multiple attribute text features separately, the average loss is taken and

denoted as L_a. The one-to-one relationship between an image and a text is transformed into a correspondence between an image and multiple texts, which helps the model learn the associations between different attributes. The Info-NCE loss is also used, and the total weighted loss L_{clip} is utilized to fine-tune the parameters of the image encoder and the text encoder. The losses are defined as:

$$L_a = \frac{1}{n}\sum_{j=1}^{n}(L_{I2T,j} + L_{T2I,j}), \tag{4}$$

$$L_{clip} = (1-\alpha)L_c + \alpha L_a, \tag{5}$$

where n is the number of attributes; $L_{I2T,j}$ is the I2T loss when constructing the text with the jth attribute; $L_{T2I,j}$ is the T2I loss when constructing the text with the jth attribute; α is the weighting parameter.

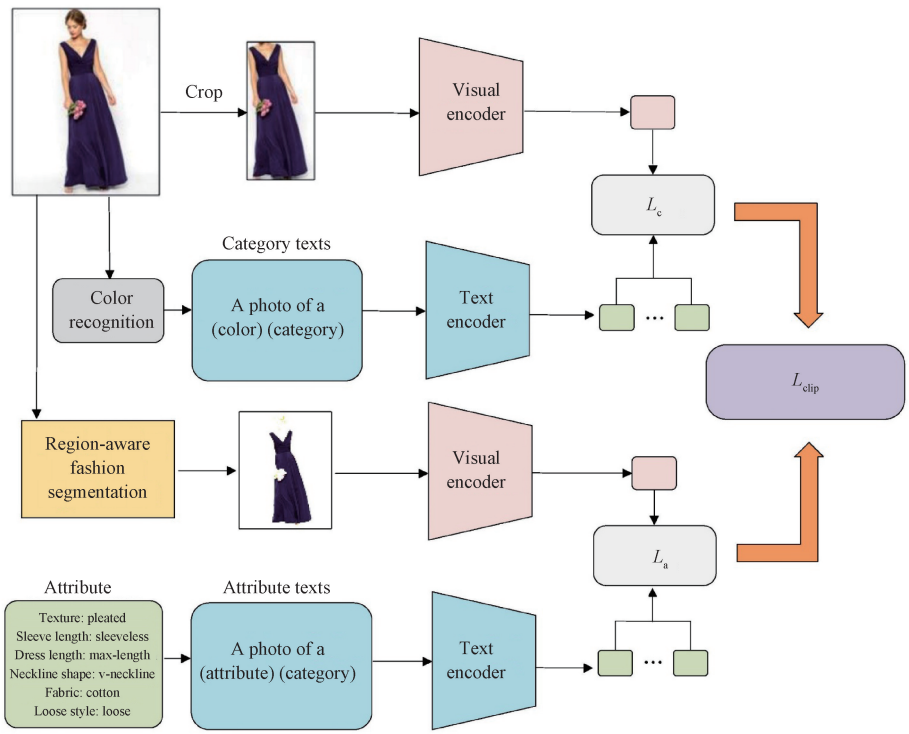

Fig. 1 Overall architecture of RaF-CLIP

1.2 Attribute guided composed image retrieval network

Composed image retrieval integrates the features of the reference image and supplementary text to form a composed query, enabling users to express their retrieval intentions more accurately. The objective function of this task can be expressed as:

$$\mathrm{M}(\boldsymbol{I}_r \oplus \boldsymbol{T}_r) \odot \boldsymbol{I}_g, \tag{6}$$

where M represents the operation of projecting the multimodal queries $\boldsymbol{I}_r$ and $\boldsymbol{T}_r$ into the same embedding space; $\oplus$ represents modifying the attributes in the reference image based on the text; $\odot$ represents similarity computation with target image features $\boldsymbol{I}_g$. Baldrati et al.[22] proposed CLIP4Cir and designed a feature addition method based on convolutional neural network (CNN). However, this method lacks effective interaction between different features. This research proposed an AGCN network using bidirectional attention based on a transformer[23] and gating mechanism. Because the textual description will modify the clothing attributes in the reference image, RaF-CLIP is used to generate features that focus on the fine-grained attributes of clothing for both images and texts. Image features and attribute text features interact separately, and then the features from the two branches are sent to their respective gating units. The gating units filter the features and control the contribution of the branch features to the final features. The flowchart of this process can be seen in Fig. 2. Q, K and V in Fig. 2 represent queries, keys and values of the self-attention layer, respectively. "Add" means residual connection, and "norm" is short for layer normalization.

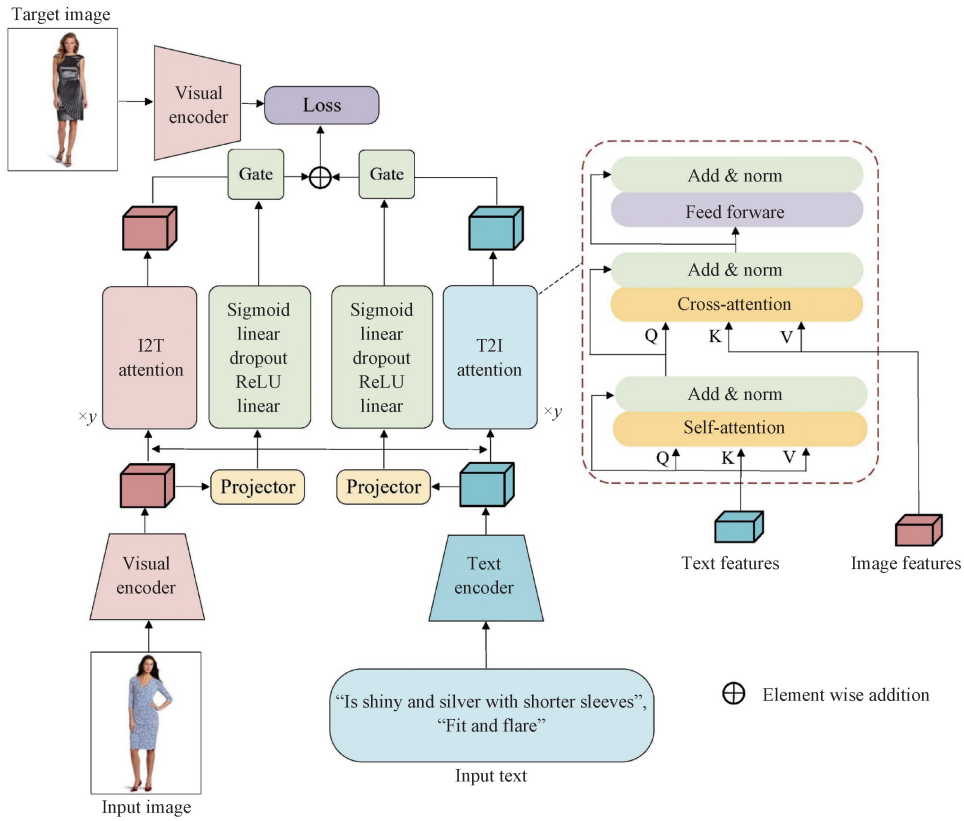

Fig. 2 Overall architecture of AGCN network

Let $\boldsymbol{I}_{\mathrm{r}} = \{\boldsymbol{i}_1, \boldsymbol{i}_2, \cdots, \boldsymbol{i}_k\}$, $\boldsymbol{T}_{\mathrm{r}} = \{\boldsymbol{t}_1, \boldsymbol{t}_2, \cdots, \boldsymbol{t}_k\}$, $\boldsymbol{I}_{\mathrm{g}} = \{\boldsymbol{f}_1, \boldsymbol{f}_2, \cdots, \boldsymbol{f}_k\}$ represent a set of reference image features, query text features, and target image features from the RaF-CLIP model, respectively. Here, $\boldsymbol{I}_{\mathrm{r}} \in \mathbf{R}^{k\times d}$, $\boldsymbol{T}_{\mathrm{r}} \in \mathbf{R}^{k\times d}$, and $\boldsymbol{I}_{\mathrm{g}} \in \mathbf{R}^{k\times d}$, where k is the number of elements in a group, and d is the feature dimension. The bidirectional attention mechanism comprises a T2I attention block and an I2T attention block, each of which is stacked y times in their respective branches. The T2I attention block first learns the contextual relationships of the text features through a self-attention layer. Then, it uses the text features as queries and the image features as keys and values, representing the process of modifying the image features based on the text features. The other branch does the opposite, representing the need to preserve parts of the image features that are not modified by the text. In the T2I attention block, the queries, keys, and values of the self-attention layer are: $\boldsymbol{Q}_{\mathrm{s}} = \boldsymbol{T}_{\mathrm{r}}\boldsymbol{W}_{Q,m}$, $\boldsymbol{K}_{\mathrm{s}} = \boldsymbol{T}_{\mathrm{r}}\boldsymbol{W}_{K,m}$, $\boldsymbol{V}_{\mathrm{s}} = \boldsymbol{T}_{\mathrm{r}}\boldsymbol{W}_{V,m}$, where $\boldsymbol{W}_{Q,m} \in \mathbf{R}^{d\times} d_{\mathrm{k}}$; $\boldsymbol{W}_{K,m} \in \mathbf{R}^{d\times d_{\mathrm{k}}}$; $\boldsymbol{W}_{V,m} \in \mathbf{R}^{d\times d_{\mathrm{v}}}$. The self-attention calculation formulas are

$$\boldsymbol{T}'_{\mathrm{r}} = \mathrm{softmax}\left(\frac{\boldsymbol{Q}_{\mathrm{s}} \cdot \boldsymbol{K}_{\mathrm{s}}^{\mathrm{T}}}{\sqrt{d_{\mathrm{k}}}}\right)\boldsymbol{V}_{\mathrm{t}}, \tag{7}$$

$$\boldsymbol{T}_{\mathrm{s}} = \mathrm{Norm}(\boldsymbol{T}_{\mathrm{r}} + \mathrm{Dropout}(\boldsymbol{T}'_{\mathrm{r}})). \tag{8}$$

After residual connection with the initial text features and the layer normalization, the features $\boldsymbol{T}_{\mathrm{s}}$ of the self-attention layer are obtained. In the cross-attention layer, the text features and image features interact within the modality. The keys and values are calculated using the image features, while the queries, keys and values are expressed as: $\boldsymbol{Q}_{\mathrm{c}} = \boldsymbol{T}_{\mathrm{s}}\boldsymbol{W}_{Q,n}$, $\boldsymbol{K}_{\mathrm{c}} = \boldsymbol{I}_{\mathrm{r}}\boldsymbol{W}_{K,n}$, $\boldsymbol{V}_{\mathrm{c}} = \boldsymbol{I}_{\mathrm{r}}\boldsymbol{W}_{V,n}$, where $\boldsymbol{W}_{Q,n} \in \mathbf{R}^{d\times d_{\mathrm{k}}}$; $\boldsymbol{W}_{K,n} \in \mathbf{R}^{d\times d_{\mathrm{k}}}$; $\boldsymbol{W}_{V,n} \in \mathbf{R}^{d\times d_{\mathrm{v}}}$; m and n represent the number of stacked attention blocks. The calculation formulas are

$$\boldsymbol{T}''_{\mathrm{r}} = \mathrm{softmax}\left(\frac{\boldsymbol{Q}_{\mathrm{c}} \cdot \boldsymbol{K}_{\mathrm{c}}^{\mathrm{T}}}{\sqrt{d_{\mathrm{k}}}}\right)\boldsymbol{V}_{\mathrm{c}}, \tag{9}$$

$$\boldsymbol{T}_{\mathrm{c}} = \mathrm{Norm}(\boldsymbol{T}_{\mathrm{s}} + \mathrm{Dropout}(\boldsymbol{T}''_{\mathrm{r}})). \tag{10}$$

Subsequently, the features $\boldsymbol{T}_{\mathrm{c}}$ pass through a feed forward network to enhance nonlinear expression ability. This layer consists of two fully connected layers and an activation function. The text branch features $\boldsymbol{F}_{\mathrm{t}}$ are obtained, and the other branch yields features $\boldsymbol{F}_{\mathrm{i}}$.

$$\boldsymbol{F}_{\mathrm{t}} = \mathrm{Norm}(\mathrm{FFN}(\boldsymbol{T}_{\mathrm{c}}) + \boldsymbol{T}_{\mathrm{c}}). \tag{11}$$

The gating mechanism utilizes two networks to

generate weights for the corresponding branches. The original image and text features pass through a projection layer, a linear layer, a ReLU activation function, a dropout, another linear layer, and a Sigmoid activation function in sequence, resulting in two weights a and $b \in (0, 1)$ for the respective branches. The final features $\boldsymbol{F}$ can be expressed as

$$\boldsymbol{F} = a \times \boldsymbol{F}_{\mathrm{i}} + b \times \boldsymbol{F}_{\mathrm{t}}. \tag{12}$$

The optimization goal of this task is to minimize the distance between the combined features $\boldsymbol{F}$ and target image features $\boldsymbol{I}_{\mathrm{g}}$, while maximizing the distance from other target images in the same batch. We use a batch-based contrastive loss L_{r} for supervision

$$L_{\mathrm{r}} = \frac{1}{B}\sum_{i=1}^{B} -\log\left(\frac{\exp\{\tau' \cdot \kappa(\boldsymbol{F}, \boldsymbol{I}_{\mathrm{g}})\}}{\sum_{j=1}^{B}\exp\{\tau' \cdot \kappa(\boldsymbol{F}, \boldsymbol{I}_{\mathrm{g}})\}}\right), \tag{13}$$

$$\kappa(\boldsymbol{F}, \boldsymbol{I}_g) = \frac{(\boldsymbol{F} \cdot \boldsymbol{I}_g)}{\|\boldsymbol{F}\|\|\boldsymbol{I}_g\|}, \tag{14}$$

where B represents the batch size; τ' is the temperature parameter controlling the scale of logits; κ represents the cosine similarity calculation.

2 Results and Discussion

2.1 Datasets and evaluation indicators

This research fine-tuned the CLIP model on the DeepFashion[3] dataset and experiments on composed image retrieval were conducted using the FashionIQ[24] dataset. A subset of the DeepFashion dataset was used for category and attribute prediction. The dataset contains 20 000 images, 50 categories and 26 sub-attributes across six attribute groups (texture, sleeve length, dress length, neckline shape, fabric and loose style). The entire dataset was divided into a training set with 16 000 images and a test set with 4 000 images. The FashionIQ dataset consisted of 77 684 fashion images, divided into training, validation and test sets, covering three different categories: dress, toptee and shirt. The training set contained 18 000 triplets, and each triplet was composed of reference image, text and target image. This research evaluated the attribute recognition and retrieval performance using the recall rate (recall@k, which was defined as the percentage of correct samples appearing in the top k retrieval results) and the precision. Additionally, the recall rate was also used to evaluate the performance of composed image retrieval.

2.2 Model training

For the RaF-CLIP model, we selected OpenCLIP[25], which was trained on the LAION-2B dataset containing 2 billion image-text pairs, as the pre-trained model. The image encoder architecture was ViT-B/32. During training, we set the batch size to 256, the learning rate to 5×10^{-7}, and used the AdamW optimizer with momentum parameter β_1 set to 0.9, β_2 set to 0.98, epsilon set to 1×10^{-6}, and weight decay set to 0.2. The α in the loss function was set to 0.5, and the model was fine-tuned for 150 epochs. For the AGCN part, we froze the RaF-CLIP parameters and only trained the AGCN network. We set the batch size to 128, the learning rate to 2×10^{-5}, and the number of epochs to 200. We adopted early stopping and mixed precision training and set the number of stacked attention blocks y to 6. PyTorch was used as the deep learning framework, and training was performed on a graphics workstation equipped with RTX 3090 GPU.

2.3 Quantitative research

The RaF-CLIP model was compared with other CLIP-based methods, including CLIP[7] and FashionCLIP[8]. The precision and the mean precision of clothing attribute recognition are shown in Table 1. As can be seen from Table 1, the RaF-CLIP model improves the mean precision by 29.42% and 6.13% compared to CLIP and FashionCLIP, respectively, demonstrating excellent clothing attribute recognition ability. In the comparison of each attribute group, except for dress length, RaF-CLIP has higher precision than CLIP and FashionCLIP. To further validate the effectiveness of the RaF-CLIP model, an image-text retrieval task was introduced to evaluate the model's performance. The recall results are shown in Table 2. Specifically, this research used the BLIP[26] model to generate text descriptions for images and used recall@k (expressed as R@k in Table 2) to evaluate the retrieval performance. Compared to CLIP, RaF-CLIP improves by 22.6% and 29.0% in the recall@50 metric for T2I retrieval and I2T retrieval, respectively, indicating that fine-tuning the CLIP model helps match images and corresponding texts in the fashion domain. To verify the effectiveness of AGCN, the performance of various models in composed image retrieval was tested on the FashionIQ validation set, and the quantitative results are shown in Table 3. In the Method column of Table 3, the phrase in front of the bracket indicates the abbreviation of the method, and the phrase inside the bracket indicates the abbreviation of the journal or conference in which the method is published. Compared to CRN (TIP 2023)[16], the recall@10 of AGCN is improved by 16.7%. Additionally, AGCN exhibits robust performance with improvements across all metrics, while CLVC-Net (SIGIR 2021)[15] experiences a decrease in performance in the average recall@10 metric.

Table 1 Quantitative results of cloth attribute recognition

Method	Precision						Mean precision
	Texture	Sleeve length	Dress length	Neckline shape	Fabric	Loose style	
CLIP	0. 553 6	0. 595 2	**0. 726 5**	0. 249 4	0. 454 5	0. 520 3	0. 512 5
FashionCLIP	0. 562 8	0. 806 3	0. 664 4	0. 506 8	0. 566 7	0. 643 1	0. 625 0
RaF-CLIP	**0. 670 2**	**0. 835 7**	0. 553 4	**0. 507 0**	**0. 737 1**	**0. 676 1**	**0. 663 3**

Table 2 Quantitative results of multimodal retrieval

Method	T2I				I2T			
	R@ 3	R@ 5	R@ 10	R@ 50	R@ 3	R@ 5	R@ 10	R@ 50
CLIP	15. 50	19. 90	27. 70	52. 55	16. 85	21. 67	29. 40	55. 65
RaF-CLIP	**23. 02**	**28. 77**	**38. 12**	**64. 42**	**26. 60**	**33. 65**	**44. 37**	**71. 82**

Table 3 Quantitative results of composed image retrieval on FashionIQ validation set

Method	Shirt		Dress		Toptee		Average	
	R@ 10	R@ 50	R@ 10	R@ 50	R@ 10	R@ 50	R@ 10	R@ 50
TIRG (CVPR 2019)[9]	18. 26	37. 89	14. 87	34. 66	19. 08	39. 62	17. 40	37. 39
VAL (CVPR 2020)[10]	22. 38	44. 15	22. 53	44. 00	27. 53	51. 68	24. 15	46. 61
CoSMo (CVPR 2021)[11]	24. 90	49. 18	25. 64	50. 30	29. 21	57. 46	26. 58	52. 31
DCNet (AAAI 2021)[12]	23. 95	47. 30	28. 95	56. 07	30. 44	58. 29	27. 78	53. 89
SAC (WACV 2022)[13]	28. 02	51. 86	26. 52	51. 01	32. 70	61. 23	29. 08	54. 70
FashionViL (ECCV 2022)[14]	25. 17	50. 39	33. 47	59. 94	34. 98	60. 79	31. 20	57. 04
CLVC-Net (SIGIR 2021)[15]	28. 75	54. 76	29. 85	56. 47	33. 50	64. 00	30. 70	58. 41
CRN (TIP 2023)[16]	30. 27	56. 97	32. 67	59. 30	37. 74	65. 94	33. 56	60. 74
AGCN	**38. 17**	**58. 88**	**35. 29**	**59. 34**	**44. 06**	**67. 02**	**39. 18**	**61. 95**

Notes: CVPR is short for conference on computer vision and pattern recognition; AAAI is short for association for the advancement of artificial intelligence; WACV is short for workshop on applications of computer vision; ECCV is short for European conference on computer vision; SIGIR is short for special interest group on information retrieval; TIP is short for transactions on image processing.

2.4 Qualitative research

Figure 3 shows the visualization results of category and attribute recognition by the RaF-CLIP model. The model accurately identifies categories, textures, neckline shapes and sleeve lengths of the upper and lower garments, with incorrect identifications marked in red. Figure 4 displays the visualization results of attribute retrieval using an image-to-image (I2I) search, where the area with a V-neck attribute is highlighted with a green box. The RaF-CLIP model is correctly retrieved, while the other two models show images with round and square neckline shapes. Figure 5 presents the visualization results of multimodal retrieval using the RaF-CLIP model, with correct results outlined in purple boxes. The model identifies new and untrained attributes in the second and the fifth groups of T2I searches, namely jeans with hole designs and orange shirts with anchor patterns. It also accurately identifies the V-neckline shape and long skirt attributes in the fifth group of I2T searches. Figure 6 shows the visualization results of composed image retrieval, where the correct target images are highlighted with green boxes. The AGCN network effectively modifies the attributes in the reference image based on the textual expression and retrieves multiple images that meet the requirements.

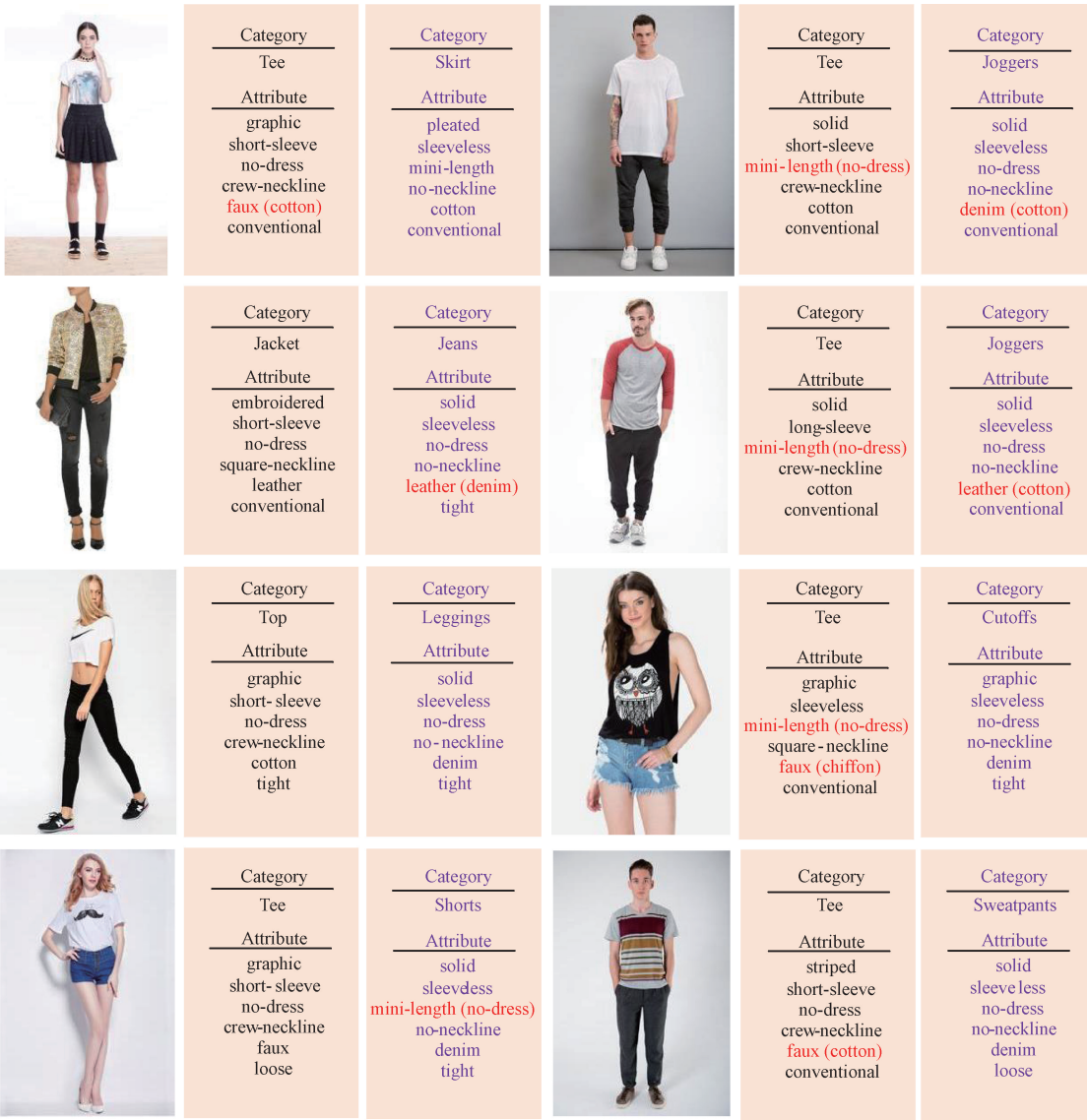

Fig. 3 Visualization results of category and attribute recognition using RaF-CLIP

Fig. 4 Visualization results of attribute retrieval

Fig. 5 Visualization results of multimodal retrieval using RaF-CLIP

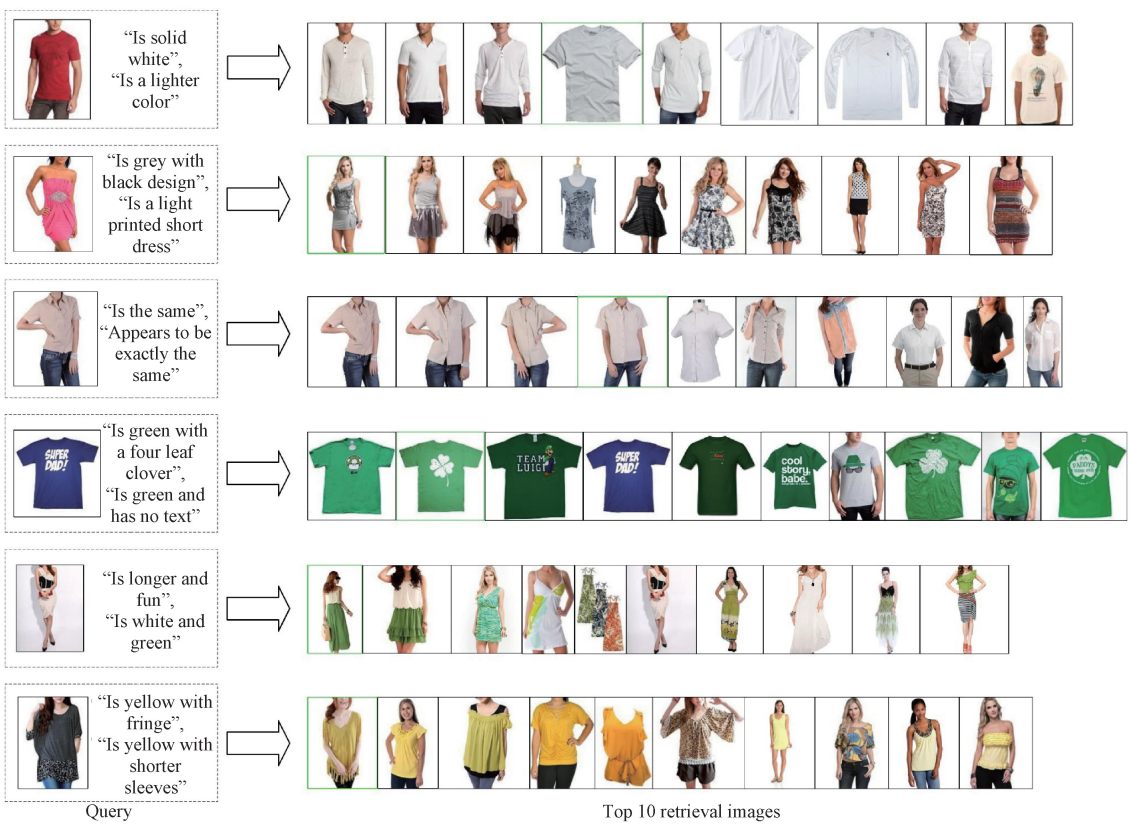

Fig. 6 Visualization results of composed image retrieval

2.5 Ablation experiments

Table 4 compares the impact of using complete images for training (Complete) versus using the RaF-CLIP architecture on the results of clothing attribute recognition. The RaF-CLIP model improves the precision by 3.5%, 6.4%, 0.3%, 2.3%, 9.0% and 0.5% for the attributes of texture, sleeve length, dress length, neckline shape, fabric and loose style, respectively, resulting in a 3.9% improvement in the mean precision. This proved that using regional-level features contributed to the recognition of fine-grained clothing attributes.

Table 5 shows the impact of the composed network structure on retrieval results. Sum represents the direct element-wise addition of features from RaF-CLIP. Only text uses only the T2I attention mechanism. Only image uses only the I2T attention mechanism. Add employs a bidirectional attention mechanism and adds the results of the two attention mechanisms with equal weight. Recall@50 for the Sum method is 31.55. Only text represents the modification process of reference image features by text features, which is an improvement over Sum, indicating that text features play a crucial role in the retrieval process. Only image represents the reference image retaining only the image features that have not been modified by text, resulting in poor retrieval performance. Add uses two attention mechanisms and performs better than the previous three groups, demonstrating the effectiveness of the bidirectional attention mechanism. The AGCN network dynamically adjusts the weights of the two branches through the gating mechanism, achieving the best results, improving 19.3% over Add in recall@10. This proves that the gating mechanism optimizes the information selection mechanism of the retrieval system, improving the performance of retrieval results and demonstrating stronger generalization capabilities.

Table 4 Impact of RaF-CLIP architecture on attribute recognition

Method	Precision						Mean precision
	Texture	Sleeve length	Dress length	Neckline shape	Fabric	Loose style	
Complete	0.647 4	0.785 3	0.551 9	0.495 7	0.676 2	0.672 3	0.638 1
RaF-CLIP	**0.670 2**	**0.835 7**	**0.553 4**	**0.507 0**	**0.737 1**	**0.676 1**	**0.663 3**

Table 5 Impact of composed network structure on retrieval results

Method	Shirt		Dress		Toptee		Average	
	R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50
Sum	18.00	32.04	11.45	28.56	17.84	34.06	15.76	31.55
Only text	24.34	42.64	21.76	42.34	27.80	49.31	24.63	44.76
Only image	7.06	17.27	4.56	13.48	6.01	15.85	5.88	15.53
Add	31.59	52.55	30.19	52.65	36.76	61.95	32.85	55.72
AGCN	**38.17**	**58.88**	**35.29**	**59.34**	**44.06**	**67.02**	**39.18**	**61.95**

3 Conclusions

Aiming at the problems that existing methods were unable to handle added attributes and capture region-level visual features, this research proposed the RaF-CLIP model for fine-grained clothing attribute recognition. By using cropped and segmented images for training, the model achieved precise alignment between clothing regions and fine-grained attribute texts. Furthermore, to realize the composed image retrieval task, this research further designed an AGCN network that selectively modified image features through attention and gating mechanisms, enabling accurate composed image retrieval. Experimental results showed that the proposed model achieved good performance in both clothing attribute recognition and composed image retrieval, providing users with the conditions to accurately express search intent and freely retrieve clothing. Future work will focus on designing efficient attention mechanisms to model the relationships between different clothing attributes, further enhancing the model's ability for attribute recognition.

References

[1] LIU J Y, LU H. Deep fashion analysis with feature map upsampling and landmark-driven attention [C]//Proceedings of the European Conference on Computer Vision (ECCV) Workshops. Berlin: Springer, 2018: 11131.

[2] LIU X, LI J, WANG J, et al. MMFashion: an open-source toolbox for visual fashion analysis [C]//Proceedings of the 29th ACM International Conference on Multimedia. New York: ACM, 2021: 3755-3758.

[3] LIU Z W, LUO P, QIU S, et al. DeepFashion: powering robust clothes recognition and retrieval with rich annotations [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2016: 1096-1104.

[4] ZENG F, ZHAO M B, ZHANG Z, et al. Joint clothes detection and attribution prediction via anchor-free framework with decoupled representation transformer [C]//Proceedings of the 31st ACM International Conference on Information & Knowledge Management. New York: ACM, 2022: 2444-2454.
[5] ZHANG S Y, SONG Z J, CAO X C, et al. Task-aware attention model for clothing attribute prediction [J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2020, 30(4): 1051-1064.
[6] LUO X, XIA D M, TAO R, et al. Fabric image retrieval based on fine-grained features [J]. *Journal of Donghua University (English Edition)*, 2024, 41(2): 115-129.
[7] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision [C]//International Conference on Machine Learning. New York: ICML, 2021: 8748-8763.
[8] CHIA P J, ATTANASIO G, BIANCHI F, et al. Contrastive language and vision learning of general fashion concepts [J]. *Scientific Reports*, 2022, 12: 18958.
[9] VO N, JIANG L, SUN C, et al. Composing text and image for image retrieval: an empirical odyssey [C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Los Alamitos: IEEE, 2019: 6432-6441.
[10] CHEN Y B, GONG S G, BAZZANI L. Image search with text feedback by visiolinguistic attention learning [C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2020: 2998-3008.
[11] LEE S, KIM D, HAN B. CoSMo: content-style modulation for image retrieval with text feedback [C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Los Alamitos: IEEE, 2021: 802-812.
[12] KIM J, YU Y, KIM H, et al. Dual compositional learning in interactive image retrieval [J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, 35(2): 1771-1779.
[13] JANDIAL S, BADJATIYA P, CHAWLA P, et al. SAC: semantic attention composition for text-conditioned image retrieval [C]//2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Los Alamitos: IEEE, 2022: 597-606.
[14] HAN X, YU L, ZHU X, et al. FashionViL: fashion-focused vision-and-language representation learning [C]//European Conference on Computer Vision. Berlin: Springer, 2022: 634-651.
[15] WEN H, SONG X, YANG X, et al. Comprehensive linguistic-visual composition network for image retrieval [C]//Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2021: 1369-1378.
[16] YANG Q, YE M, CAI Z, et al. Composed image retrieval via cross relation network with hierarchical aggregation transformer [J]. *IEEE Transactions on Image Processing*, 2023, 32: 4543-4554.
[17] ZHA J H, YAN C R, ZHANG Y T, et al. Image retrieval with text manipulation by local feature modification [J]. *Journal of Donghua University (English Edition)*, 2023, 40(4): 404-409.
[18] LI F, PAN H S, SHENG S X, et al. Image retrieval based on vision transformer and masked learning [J]. *Journal of Donghua University (English Edition)*, 2023, 40(5): 539-547.
[19] LIANG J H, LIU Y, VLASSOV V. The impact of background removal on performance of neural networks for fashion image classification and segmentation [EB/OL]. (2023-8-18) [2024-5-7]. https://doi.org/10.1109/CSCE60160.2023.00323.
[20] HE K M, FAN H Q, WU Y X, et al. Momentum contrast for unsupervised visual representation learning [C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2020: 9726-9735.
[21] KIRILLOV A, MINTUN E, RAVI N, et al. Segment anything [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2023: 4015-4026.
[22] BALDRATI A, BERTINI M, URICCHIO T, et al. Composed image retrieval using contrastive learning and task-oriented clip-based features [J]. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2024, 20(3): 1-24.
[23] ASHISH V, NOAM S, NIKI P, et al. Attention is all you need [C]//31st Annual Conference on Neural Information Processing Systems. San Diego: NIPS, 2017, 30.
[24] WU H, GAO Y P, GUO X X, et al. FashionIQ: a new dataset towards retrieving images by natural language feedback [C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2021: 11302-11312.
[25] CHERTI M, BEAUMONT R, WIGHTMAN R, et al. Reproducible scaling laws for contrastive language-image learning [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2023:

2818-2829.
[26] LI J, LI D. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation[C]//International conference on machine learning. New York: ICML, 2022: 12888-12900.

面向属性识别和组合检索的区域感知时尚对比学习

王康平，赵鸣博*
东华大学 信息科学与技术学院，上海 201620

摘　要：服装属性识别已成为一项关键技术，使用户能够自动识别服装的特征，并搜索具有相似属性的服装图片。然而，现有方法无法识别新添加的属性，并且可能无法捕获区域级别视觉特征。为解决上述问题，该研究提出一种区域感知时尚对比语言图像预训练（region-aware fashion contrastive language-image pre-training，RaF-CLIP）模型。该模型将裁剪和分割的图像与类别和多个细粒度属性文本进行对齐，通过对比学习实现时尚区域与相应文本的匹配。服装检索基于用户指定的服装类别和属性来找到合适的服装，为进一步提高检索的准确性，该研究在 RaF-CLIP 模型上引入属性引导的组合网络（attribute-guided composed network，AGCN），并将其作为附加组件，专用于组合图像检索任务。该任务旨在根据文本表达修改参考图像以检索预期的目标。通过采用基于 transformer 的双向注意力和门控机制，该网络实现了图像特征和属性文本特征的融合与选择。试验结果表明，所提出的模型在属性识别任务中平均精度达到 0.663 3，在组合图像检索任务中 recall@10（recall@k 表示正确样本出现在前 k 个检索结果中的百分比）指标达到 39.18，满足用户通过图像和文本自由搜索服装的需求。

关键词：属性识别；图像检索；对比语言图像预训练（CLIP）；图像文本匹配；transformer