Complex & Intelligent Systems (2024) 10:6103–6116
https://doi.org/10.1007/s40747-024-01456-6

ORIGINAL ARTICLE

Residual serialized cross grouping transformer for small scale sketch face recognition

Kangning Du[1,2] · Yinkai Wang[1,2] · Jianqiang Yin[1,2] · Lin Cao[1,2] · Yanan Guo[1,2]

Received: 7 December 2023 / Accepted: 20 April 2024 / Published online: 31 May 2024

Abstract
Sketch face recognition has recently gained significant attention in the field of computer vision due to its ability to quickly identify matched pairs of optical and sketch images. This technology has the potential to greatly improve the efficiency of law enforcement agencies in criminal investigations. However, there are still challenges that need to be addressed in sketch face recognition algorithms, such as modal differences and limited sample sizes. To overcome these issues, this study proposes a Residual Serialized Cross Grouping Transformer (RSCGT), which contains a residual serialized module to reduce the computation complexity, a two-layer Cross Grouping Transformer module that is capable of extracting modality-invariant context features, a domain adaptive module to mitigate the impact of modal differences. Additionally, we introduce a meta-learning training strategy to augment the generalization ability of this model. Experimental results demonstrate that the RSCGT achieves high accuracy in sketch face recognition tasks, even with small-scale datasets.

Keywords Sketch face recognition · Small sample training · Residual serialized · Cross Grouping Transfromer

Introduction

Sketching is an artistic approach that utilizes fundamental lines and shading to depict the structure, contours, and textures of a photograph. In the pursuit of enhancing the efficacy of law enforcement, the sketch face recognition algorithm has emerged as a powerful tool for associating optical images with their corresponding sketches. This algorithm has extensive and enduring practical applications within the field of criminal investigations, enabling effective identification and tracking of criminals, ultimately contributing to the preservation of social order and stability.

However, sketch face recognition algorithms still face challenges such as modality differences and small sample problems. Therefore, the research on high-accuracy sketch face recognition algorithms holds significant practical importance. Firstly, optical images are captured using optical devices and depict real scenes, providing a high degree of realism and detail. On the other hand, sketch images are manually drawn, exhibiting stronger expressiveness and artistic qualities. These divergent representations give rise to modality differences between optical and sketch images, which greatly impact the performance of sketched image recognition algorithms. Secondly, the sketched face datasets utilized for algorithm research are often limited in size, making the models susceptible to overfitting. Thus, the objective of this study is to enhance the extraction of modality invariant features between optical and sketch images, while also improving the generalizability of the model.

To tackle the aforementioned challenges, numerous methods for sketch face recognition have been proposed. Traditionally, manual feature design has been the dominant approach in this field. However, in recent years, deep learn-

✉ Lin Cao
charlin@bistu.edu.cn

Kangning Du
kangningdu@bistu.edu.cn

Yinkai Wang
yinkaiwang@bistu.edu.cn

Jianqiang Yin
2395986951@qq.com

Yanan Guo
yananguo@bistu.edu.cn

1 Key Laboratory of the Ministry of Education for Optoelectronic Measurement Technology and Instrument, Beijing Information Science and Technology University, Beijing 100101, China

2 School of Information and Communication Engineering, Beijing Information Science and Technology University, Beijing 100101, China

ing has gained popularity due to its remarkable performance. In particular, Convolutional Neural Networks (CNNs) have been widely adopted as the backbone network for feature extraction. These networks can automatically learn and extract cross-modal consistent information, which is crucial for accurate recognition. However, CNNs primarily function as local feature extractors, posing difficulties in capturing long-distance features and establishing global connections among them. Consequently, extracting modality-invariant semantic features becomes a challenging task.

Transformer [1], which utilizes a self-attention mechanism, introduces a novel approach to address these limitations. Unlike CNNs, Transformers have the ability to attend to various positions within the input sequence during the processing of each input. This enables them to capture dependencies between different positions effectively, enhancing the model's understanding of the overall context. By employing this self-attention mechanism, Transformers can overcome the constraints of CNNs in extracting solely local features. Instead, they excel at capturing global connections among features. As a result, Transformers can learn comprehensive contextual relationships and emphasize the significance of discriminative features in recognition tasks.

It should be noted that despite the advantages of Transformer architectures, such as the Vision Transformer (ViT) [2], their computational complexity is often high. This can present challenges when dealing with small-scale sketch-face datasets, where computational and resource constraints may make the direct application of Transformers unsuitable. Fortunately, a recent variant called Swin-Transformer [3] has been developed to address these concerns. The Swin-Transformer improves performance and adaptability on smaller datasets by introducing certain modifications. Specifically, it restricts self-attention to local regions and establishes interactive connections between different regions to effectively capture global connections. By adopting this approach, the Swin-Transformer avoids the quadratic complexity of the original Transformer model. It reduces the number of tokens involved, enhancing performance and improving adaptability on smaller samples.

However, it is important to note that these methods introduce a certain level of redundancy when trying to obtain global connections through local interactions. The two tokens responsible for performing local self-attention calculations may perform duplicate calculations, resulting in unnecessary computational consumption. To address this issue, hybrid models such as BoTNet [4] and Axial-ResNet [5] have been developed. These hybrid models integrate the strengths of both CNNs and Transformers. By effectively leveraging CNN's proficiency in handling high-dimensional image data and the Transformer's capability to capture global information from features, these hybrid models offer a more efficient solution. By combining these two techniques, these hybrid models overcome the heightened complexity associated with directly applying Transformer models to high-resolution image embedding tasks. Consequently, in sketch face recognition scenarios, hybrid models prove to be more suitable and efficient options to consider.

To address the issue of excessive complexity in Transformer models for sketch face recognition, this paper proposes a Residual Serialized Cross Grouping Transformer (RSCGT), which contains a residual serialized module to reduce the computational cost caused by high-dimensional image embedding in Transformers, a two-layer Cross Grouping Transformer module to effectively captures contextual connections among features, and a domain adaptive module to handle the modality differences between sketch images and optical images. In addition, we adopt an effective meta-learning training strategy specifically designed for small samples, preventing overfitting issues. The main contributions of our work can be summarized as follows:

1. We propose a novel residual serialized module that reduces the computational cost of Transformer models by efficiently embedding high-dimensional image features.
2. We propose a computationally efficient Cross Grouping Transformer module that captures contextual relationships of features. The module divides the complex self-attention process into two stages: grouping multi-head self-attention and cross-group multi-head self-attention. This approach enables the representation of global information in lower-complexity features.
3. We adopt a domain adaptive module to address the modality differences between sketch images and optical images. Our module mitigates the impact of modal differences on recognition performance, thereby enabling our method to be applied to both sketch and optical images for more comprehensive and accurate face recognition.
4. We conducted comprehensive experiments on three sketch face recognition datasets, including UoM-SGFS, CUFSF, and PRIP-VSGC, to evaluate the accuracy of our approach. The experimental results demonstrate consistent and significant improvements compared to mainstream methods and indicate the efficacy of our method in handling small sample data.

Related works

Sketch face recognition methods

Sketch face recognition methods have witnessed a transition from traditional hand-designed approaches to deep learning methods due to the urgent demands and extensive applications in the field of criminal investigation. With the

introduction of new algorithms that yield improved results, researchers have continuously pursued their research in this area.

The traditional method of sketch face recognition primarily relies on manual feature design to accomplish the task. Klare et al. [6] introduced the Local Feature-based Discriminant Analysis (LFDA) approach, which utilizes Scale-Invariant Feature Transform (SIFT) descriptors [7] and Multiscale Local Binary Patterns (MLBPs) [8] to represent both sketch and optical images. Han et al. [9] introduced Component-based Representation (CBR), which employs an Active Shape Model (ASM) to automatically detect facial landmarks and represent facial components using MLBPs. The similarities of the features of each component are then fused to match optical and sketch images. Bonnen et al. [10] proposed a component-based framework for facial alignment and representation. This method employs ASM to identify the location of facial landmarks, aligns components using Procrustes analysis [11], and represents components using MLBPs to achieve component-based facial alignment and representation. However, these methods have limited representational power in describing the highly nonlinear relationships between cross-modal images and face challenges in improving recognition rates. Furthermore, these methods cannot automatically learn and extract modality-invariant features like CNNs.

With the rapid progress of deep learning, the effectiveness of deep learning-based sketch face recognition methods has surpassed that of traditional hand-designed approaches, establishing itself as the dominant approach in this field. Currently, deep learning methods for sketch face recognition can be categorized into two types: intra-modal methods and inter-modal methods.

Intra-modal methods refer to approaches that reduce modality differences by transforming cross-modal images into the target modality, followed by identity recognition using a recognition model. Zhang et al. [12] designed an end-to-end Fully Convolutional Network (FCN) to learn the mapping relationship between optical images and sketch images. By inference and learning, their method can generate sketch images corresponding to optical images. While intra-modal methods provide an intuitive way to obtain images in the target modality, their performance heavily relies on the quality of the synthesized images [13]. Moreover, intra-modal methods face difficulties in capturing the nonlinear relationship of modality transformations when the modality discrepancy is significant.

The inter-modal methods reduce modality differences by mapping cross-modal features to a common subspace. These methods focus on learning classifiers that maximize inter-class differences and minimize intra-class differences, accomplishing identity recognition by extracting modality invariant features. Wan et al. [14] proposed a sketch-based face recognition method based on transfer learning. They designed a three-channel CNN structure and used triplet loss to learn discriminative features and reduce intra-class differences. They also introduced a hard triplet sample selection strategy to increase the number of training samples and accelerate model convergence. Gui et al. [15] proposed a multi-modal recognition method, which leverages feature-level knowledge distillation to achieve complementary advantages between Transformers and CNNs, enhancing feature extraction capabilities. Cheraghi et al. [16] proposed a coupled Sketch-Photo Net (SP-Net) that consists of two branches, S-Net and P-Net, to learn discriminative features between sketch and photo images. This method also utilized contrast loss to discover coherent visual structures between sketch and photo images. Guo et al. [17] proposed a deep metric learning method based on Domain Alignment Embedding Network (DAEN). They designed a meta-learning training set strategy to alleviate overfitting caused by small samples and introduced a domain alignment embedding loss to guide the feature embedding network in learning discriminative features. However, these methods struggle to extract discriminative features between cross-modal images when the modality gap is large. Moreover, these CNN-based methods also find it challenging to eliminate the negative impact of semantic errors.

Vision transformer methods

Ever since the inception of the Transformer model by the Google team in 2017, it has become a seminal and highly popular paradigm in the realm of the natural language processing (NLP) field. With its self-attention mechanism, Transformer has achieved remarkable accomplishments, garnering considerable attention from researchers. This has also led experts in Computer Vision (CV) to explore its applicability in the visual domain due to its exceptional modeling capabilities. Transformer has witnessed rapid advancements in the CV field, exhibiting superior performance compared to CNN-based models in various CV tasks, especially when there is substantial and high-quality data support. The Vision Transformer (ViT) was proposed by Dosovitskiy et al for image recognition which processes image patch sequences by employing stacked Transformer encoders. The encoded sequences are then passed through a classification head for target classification. Although ViT has high complexity, it achieves excellent results when trained with large datasets compared to state-of-the-art (SOTA) CNNs. Chakravarthi et al. [18] effectively extract emotional features from EEG signals and model the changes of these features over time by combining CNN and LSTM, achieving accurate recognition of emotions. Chen et al. [19] proposed the image Processing Transformer (IPT) for low-level computer vision tasks. IPT employs contrastive learning for model training and opti-

mization, leading to promising results in various benchmark tests. The outstanding performance of Transformers in various tasks has led researchers from different fields to conduct in-depth research on Transformers.

The outstanding performance of Transformers in various tasks has prompted researchers from different fields to conduct in-depth research on Transformers. The Transformer model is a type of model based on a self-attention mechanism, which can establish global dependencies within a sequence. However, unlike textual sequences, images are two-dimensional structures and do not have a clear sequential relationship. Therefore, we need an efficient method to serialize images into one-dimensional sequences to reduce training costs.

In ViT, a commonly used method for image patching and serialization was proposed. This method involves dividing the feature map of an image into smaller image patches, where the pixel features within each patch are concatenated in the dimensionality direction and mapped into visual tokens. The resulting sequence obtained from this method has a high number of tokens and dimensions. However, since Transformers are sequence-to-sequence models, their parameter count and computational complexity are closely related to the number and dimensions of tokens. Therefore, this serialization method is not suitable for the serialization of high-resolution images as it would lead to a significant increase in the number of tokens and computational demands. Ying et al. [20] proposed a Graph Transformer that learns an image's node or graph representation to incorporate graph structure information into the Transformer. However, due to the lack of spatial order in graph nodes and the high computational complexity of self-attention computation in Transformers, graph convolution Transformers are not adaptable to large graphs with many nodes. Wu et al. [21] proposed Visual Transformers (VT), which introduced a new method for image serialization. Visual Transformers utilize tokens to represent image feature maps as a variable number of visual semantic labels. By adjusting the length of the embedding sequence in the Transformer, the high computational complexity of token modeling can be avoided. The VT module is responsible for token generation, modeling, and reshaping processes.

Proposed method

To address the aforementioned issues, in this section, we propose a novel sketch face recognition method (RSCGT) based on the Transformer, which addresses the issue of adaptability of Transformer models in the scenario of small sample sizes of sketch face data. Instead of serializing images and inputting them into the feature extraction method of the Transformer model, we introduce a grouping strategy within the Transformer to reduce the computational complexity of the Transformer while maintaining global connections. We first provide a brief introduction to our framework in Sect. Overview, and then give a detailed description of our model in Sect. Meta-learning training strategy to Sect. Domain adaptive module.

Algorithm 1 RSCGT model

Input: input Sample for meta-learning task $Q_t = \{P_t, S_t\} = \{p_1, p_2, \ldots, p_k, s_1, s_2, \ldots, s_k\}$
Output: output result

```
Initialize optimizer, iter T.
for i = 1 : T do
    Extract image feature X_in (X_in ∈ Q_t)
    Serialized to Sequence T = R(X_in)
    Position Encoding T_in = T + PE(T)
    Grouped Sequence T_i = G(T_in, y)
    Build Grouping Transformer
    Cross-Grouped Sequence T_i to T_i'
    Build Cross Grouping Transformer
    Loss.backward()
end for
return result
```

Overview

Given an input sketch image X, our goal is to extract the feature and Serialized by Residual Serialized module and input the Sequence T to Cross Grouping Transformer module to correctly recognize the sketch.

The proposed RSCGT model is presented in Fig. 1. Firstly, we introduce a Residual Serialized module to extract local features of images. Then, we proposed novel a two-layer Cross Grouping Transformer module to capture the global connection of features and prevent excessive complexity in Transformer models. Moreover, we proposed the improved Feed Forward Neural Network (FFN) to balance the computational cost between the self-attention layer and the FFN [22], as well as to further reduce the number of parameters in the model. To effectively train RSCGT and prevent model overfitting, we employ a meta-learning training strategy [17], which is especially efficient for small-sample scenarios.

We represent the RSCGT as $f(\cdot|w)$, where w is the model parameter. The optical image p_i or the sketch image s_i in the meta-task batch sample is passed through $f(\cdot|w)$ to obtain the corresponding feature vector $f(p_i|w)$ or $f(s_i|w)$, respectively. we optimize the feature extraction network by introducing domain alignment embedding loss [17] to continuously reduce the distance between $f(p_i/w)$ and $f(s_i/w)$ with the same label in space.

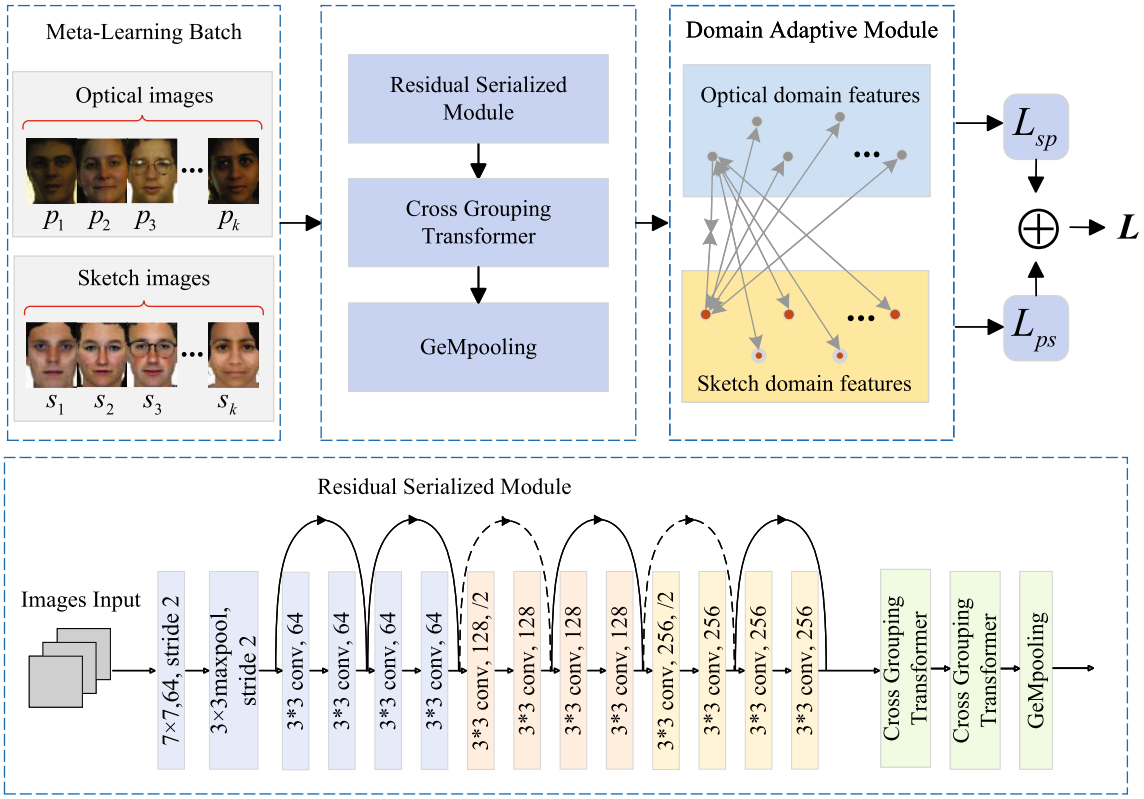

Fig. 1 RSCGT. We propose a novel transformer (RSCGT), which includes residual serialized modules, a two-layer Cross Grouping Transformer module, and a domain adaptation module

Meta-learning training strategy

To improve the generalization ability and the capability of small sample training of the model, we adopt a meta-learning training strategy that is particularly effective for our model. Given a training set $D_{train} = \{p_1, p_2, \ldots, p_N, s_1, s_2, \ldots, s_N\}$, where $P = \{p_i\}_{i=1}^{N}$ represents optical images and $S = \{s_i\}_{i=1}^{N}$ represents sketch images, i is the label of the image, and there are a total of N pairs of training samples. The meta-learning training strategy divides the entire model's training process into a series of small meta-learning tasks. The model achieves the final sketch facial recognition task by completing these meta-learning tasks. In each meta-learning task, the strategy first randomly selects $K(K < N)$ pairs of sketch and photo images from the training set D_{train} and resets the labels of the selected samples to $\{1, 2, ..., K\}$. The query set Q_t for the meta-learning task is formed by the selected samples: $Q_t = \{P_t, S_t\} = \{p_1, p_2, \ldots, p_k, s_1, s_2, \ldots, s_k\}_{t=1}^{T}$. In Q_t, P_t serves as the support set for photo images, S_t serves as the support set for sketch images, and $t = \{1, 2, ..., T\}$ indicates the number of times the meta-learning task is executed.

Residual serialized module

For the CNN backbone in the hybrid model, we adopted a residual network that supports model depth expansion, following the architecture of BoTNet [23]. The Transformer module in our model consists of Multi-Head Self-Attention (MHSA) and an improved Feed Forward Neural Network (FFN), which contributes to a deeper network depth. ResNet18 has achieved excellent performance in object detection and face recognition tasks, so we utilized the first four stages of ResNet18 as the serialized module of RSCGT. Additionally, we adopt a two-layer Cross Grouping Transformer module and GeMpooling, while retaining skip connections to support network depth expansion and accelerate model training.

The image sample $x(x \in Q_t)$ in the meta-learning task is pre-trained to a size of 256×256, and the feature map size changes after downsampling by the residual network are shown in the output in Fig. 1. The input image is subjected to the first four stages of CNN outputting low-resolution feature maps:

$$X_{in} = CNNs(x) \tag{1}$$

Transformer is a sequence-to-sequence model where we serialize the low-resolution feature map pixel-by-pixel into token sequences embedding in the Cross Grouping Transformer:

$$T = R(X_{in}) \tag{2}$$

where $R(\cdot)$ is the reshaping operation and the feature map $X_{in} \in R^{d\times H\times W}$ is transformed into a visual token sequences $T \in R^{HW\times d}$ by the reshaping operation. The absolute position encoding sequence [2] provides positional information about the token in space. The embedding token sequences are obtained by adding the positional encoding to the token sequences as follows:

$$T_{in} = T + PE(T) \tag{3}$$

The embedding tokens T_{in} capture the contextual relationships of the entire sequence through the grouping MHSA and cross-group MHSA of the Cross Grouping Transformer, which includes discriminative information between cross-modal images. The modeled features are then filtered for redundant information using GeMpooling operation [24], and subsequently mapped to a common space through the domain adaptation module to reduce the modality differences between cross-modal image features.

Cross grouping transformer module

Transformer consists of two main modules: the Multi-Head Self-Attention module and the FFN module. Transformer utilizes MHSA to calculate global relationships between embedding tokens. Due to the fact that MHSA performs self-attention calculations within the global token scope, it incurs a computational complexity that is quadratic in relation to the number of tokens. To alleviate this issue, we propose a new Transformer module called Cross Grouping Transformer, as illustrated in Fig. 2.

The Cross Grouping Transformer consists of the grouping MHSA module, the cross-group MHSA module, and the improved FFN module. In Cross Grouping Transformer, the grouping MHSA avoids the global self-attention computation on all tokens by grouping the tokens. Each group of tokens performs self-attention computation within the group, reducing the computational complexity to a linear relationship. The interaction between different groups is achieved through the cross-group MHSA. It reorganizes the tokens with the same relative position within the group in the grouping MHSA into a new group to perform cross-group MHSA computation. This enables the Cross Grouping Transformer to perform global self-attention computation on all tokens.

We define the form of grouping as a grouping strategy that satisfies the condition where the number of tokens N($N = H \times W$) is equal to the product of the number of groups and the number of tokens within each group (number of groups and the number of tokens within each group are greater than 1). Assuming that the token sequences are grouped according to $N = x \times y$ in grouping MHSA, where x represents the number of groups and y represents the number of tokens within each group. First, we input the sequence T in to Grouping Transformer and calculate the MHSA within each group, with the grouping strategy, we only consider self-attention in length y, resulting an in-group computation complexity $O(yNd)$. Based on the regrouping mechanism of cross-group MHSA, we know that the number of groups and the number of tokens within each group are y and x, respectively. The grouping strategy can be represented as $N = y \times x$. We then calculate the cross-group self-attention in length x, and the computational complexity of its self-attention is $O(xNd)$. Thereby, we transform the global multi-head self-attention computation into two local self-attentions with computation complexity $O((x+y)Nd)$, which is lower than the complexity of $O(xyNd)$ in the original Transformer. Consequently, the Cross Grouping Transformer can capture the global connections between embedding tokens with lower complexity.

(1) *Grouping Multi-Head Self-Attention Mechanism* The grouping MHSA mechanism is illustrated in Fig. 2, where the sequence of tokens with location encoding is divided into x groups of token sequences with y tokens using the grouping strategy $N = x \times y$:

$$T_i = G(T_{in}, y) \in \{T_1, T_2, \ldots, T_x\} \tag{4}$$

where $G(\cdot, y)$ represents the grouping function that groups the token sequences according to the number of tokens y. The grouped token sequences $T_i \in R^{y\times d} s.t. i \in \{1, 2, \ldots, x\}$, represent x groups of token sequences that complete the MHSA computation in parallel. This computation captures the contextual connections between the tokens within the groups. The groups of tokens linearly map the embedding token sequences to the query matrix, key matrix, and value matrix using the transfer matrices $W_i^q, W_i^k, W_i^v \in R^{d\times d/h}$:

$$Q_{ij}, K_{ij}, V_{ij} = T_i W_j^q, T_i W_j^k, T_i W_j^v, \quad j \in \{1, 2, \ldots, h\} \tag{5}$$

The query matrix $Q_{ij} \in R^{y\times d_q}$, key matrix $K_{ij} \in R^{y\times d_k}$, and value matrix $V_{ij} \in R^{y\times d_v}$ are defined, where h is the number of heads, with $d_q, d_k, d_v = d/h$. The self-attention matrix between intra-group tokens is obtained by performing scaled dot-product attention on Q_{ij} and K_{ij}. The self-attention matrix is normalized using the $softmax(\cdot)$ and then multiplied with V_{ij}, resulting in the single-head self-

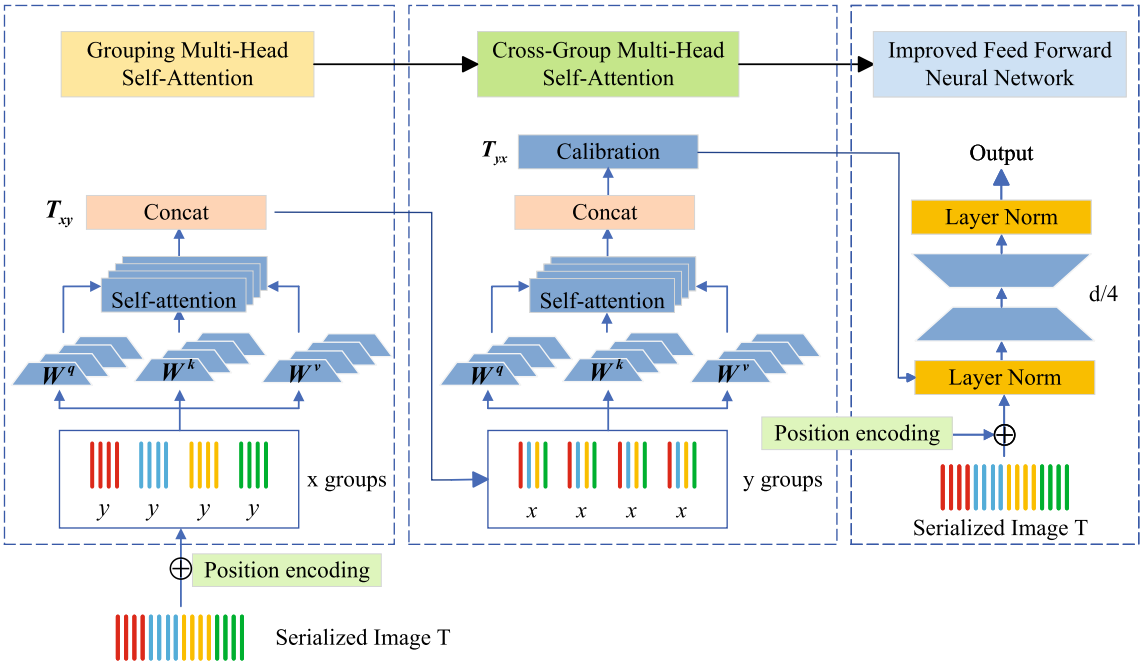

Fig. 2 Cross Grouping Transformer. We propose the Cross Grouping Transformer model, which reduces computational complexity by dividing the token sequences into multiple groups and calculating multi-head self-attention through grouping strategies

attention modeling of intra-group tokens:

$$head_{ij} = softmax\left(\frac{Q_{ij}K_{ij}^{T}}{\sqrt{d_k}}\right)V_{ij} \tag{6}$$

where $head_{ij} \in R^{y\times d/h}$ represents the modeling results of the j-th head of the i-th group in the grouping self-attention. The modeling results of all single-head self-attention in each group are concatenated in the token dimension to obtain the MHSA modeling results of the group token sequences T_i:

$$GMSA(T_i) = Concat(head_{i1}, head_{i2}, \ldots, head_{ih}) \tag{7}$$

where $GMSA(T_i) \in R^{HW\times d/h}$ represents the MHSA modeling results of the i-th group of token sequences. By concatenating the results of modeling self-attention on all individual heads within each group in the direction of the token dimension, we obtain the MHSA modeling results for token sequences of the group T_i as follows:

$$T_{xy} = Concat(GMSA(T_1), GMSA(T_2), \ldots, GMSA(T_x)) \tag{8}$$

where T_{xy} has contextual connections between tokens within the same group, but lacks self-attention interaction between tokens from different groups, preventing the establishment of contextual connections between different groups and global features.

(2) *Cross-Group Multi-Head Self-Attention Mechanism*

To overcome the shortcomings of grouping self-attention, the Cross Grouping Transformer regroups the in-group modeling tokens of the grouping MHSA output by combining them in a cross-group T_{xy}. Therefore, cross-group interactions between the original grouped tokens are achieved to capture the global connections among all tokens. The cross-group MHSA is shown in Fig. 2. According to the characteristics of grouping and cross-group, the cross-group will follow the grouping strategy of $N = y \times x$. The grouping results in a sequence of tokens with x tokens in group y:

$$T_i^{'} = G(T_{xy}, x) \in \{T_1^{'}, T_2^{'}, \ldots, T_x^{'}\} \tag{9}$$

where grouping token sequences $T_i^{'} \in R^{x\times d} s.t.i \in \{1, 2, \ldots, y\}$, y grouped token sequences perform the MHSA computation concurrently and the token sequences $T_i^{'}$ are mapped again to the new query matrix, key matrix, and value matrix by W_i^q, W_i^k, W_i^v in grouping MHSA:

$$Q_{ij}^{'}, K_{ij}^{'}, V_{ij}^{'} = T_i^{'}W_j^q, T_i^{'}W_j^k, T_i^{'}W_j^v, \quad j \in \{1, 2, \ldots, h\} \tag{10}$$

where query matrix $Q_{ij}^{'} \in R^{x\times d_q}$, the key matrix $K_{ij}^{'} \in R^{x\times d_k}$ and the value matrix $V_{ij}^{'} \in R^{x\times d_v}$. The transfer matrix performs parameter sharing in both grouping MHSA and cross-group MHSA to reduce the parameter pressure on the model. $Q_{ij}^{'}, K_{ij}^{'}, V_{ij}^{'}$ performs self-attention calculations to

obtain the modeling results for single-headed self-attention within the cross-group MHSA:

$$head_{ij}^{'} = softmax\left(\frac{Q_{ij}^{'} K_{ij}^{'T}}{\sqrt{d_k}}\right) V_{ij}^{'} \tag{11}$$

All single-headed self-attention modeling results within the group are concatenated in the direction of the token dimension to obtain the MHSA modeling results within the group for the token sequences $T_i^{'}$:

$$GMAS(T_i^{'}) = Concat(head_{i1}^{'}, head_{i2}^{'}, \ldots, head_{ij}^{'}) \tag{12}$$

The results of MHSA modeling of all groups are concatenated in the direction of the number of tokens, and the position of tokens in space is changed due to tokens reorganization in cross-group MHSA modeling, so the concatenated tokens should be reset by calibration. The calibrated token sequences are then mapped by the parameter matrix $W^0 \in R^{d\times d}$ to obtain the cross-group MHSA modeling results:

$$T_{yx} = Correct(Concat(GMSA(T_1^{'}), GMSA(T_2^{'}), \ldots, GMSA(T_y^{'})))W^o \tag{13}$$

where $Correct(\cdot)$ is a calibration function to correct the misaligned token sequences. Summing $T_{yx} \in R^{HW\times d}$ and embedding tokens to complete the jump connection. The modeling output of the Cross Grouping Transformer self-attentive layer is then obtained after layer normalization as follows:

$$T^{'} = LayerNormalization(T_{yx} + T_{in}) \tag{14}$$

As a result, the global information between the visual tokens is available in $T^{'}$ obtained after grouping MHSA and cross-group MHSA modeling.

(3) *Improved Feed Forward Neural Network*

The Feed Forward Neural Network in Transformer consists of two linear layers. It first expands the dimension of the input tokens by a factor of 4 and then reduces it back to the original dimension by a factor of 4. When the dimension of the tokens is large, the FFN consumes a significant amount of computational resources. This is detrimental to the global context modeling in Transformer, as the FFN does not participate in the computation of global context and only maps the modeling results from the self-attention layer. In order to balance the computational cost between the self-attention layer and the FFN, and further reduce the parameter count of the model, the improved FFN in the Cross Grouping Transformer first reduces the dimension of the tokens by a factor of 4 and then expands it by a factor of 4. The structure of the improved FFN is shown in Fig. 2. The output of $T^{'}$ after the FFN mapping is obtained.

$$FFN(T^{'}) = \sigma(T^{'}F_1)F_2 \tag{15}$$

where $F_1 \in R^{d\times d/4}$, $F_2 \in R^{d/4\times d}$ is the weight matrix and $\sigma(\cdot)$ is the Relu activation function. $FFN(T^{'})$ is added with the modeling output of the self-attentive layer to complete the jump connection and perform layer normalization. Finally, it is added with the embedding token T_{in} to obtain the encoded output of Cross Grouping Transformer:

$$T_{out} = LayerNormalization(T^{'} + FFN(T^{'})) + T_{in} \tag{16}$$

Domain adaptive module

In meta-learning tasks, K pairs of samples are randomly selected as batch samples. The set of these K pairs of samples is the query set of the meta-task, denoted as $Q_t = \{p_1, p_2, \ldots, p_k, s_1, s_2, \ldots, s_k\}$. The optical images in Q_t are used as the optical image support set $P_t = \{p_1, p_2, \ldots, p_k\}$, and the sketch images in Q_t are used as the sketch image support set $S_t = \{s_1, s_2, \ldots, s_k\}$. The optical images $p_i (i \in \{1, 2, \ldots, k\})$ and sketch images $s_i (i \in \{1, 2, \ldots, k\})$ in Q_t are extracted using RSCGT to obtain the optical image feature vectors $f(p_i/w)$ and sketch image feature vectors $f(s_i/w)$. The domain alignment embedding loss uses Euclidean distance to measure the distance between features. The distance between the features of optical images in the query set Q_t and the features of sketch images in S_t, as well as the distance between the features of sketch images in the query set Q_t and the features of optical images in P_t, can be represented by the Euclidean distance between feature vectors:

$$d(p_i, s_i) = ||f(p_i/w) - f(s_i/w)|| \tag{17}$$

$$d(s_i, p_i) = ||f(s_i/w) - f(p_i/w)|| \tag{18}$$

where $||\cdot||$ represents the Euclidean distance, the domain alignment embedding loss increases the similarity between images of the same label by reducing the negative Euclidean distance on cross-domain features. For an optical image p_i or a sketch image s_i, the labels can be predicted by the $softmax(\cdot)$ function on the negative Euclidean distance between them and all the cross-domain images in the meta-task:

$$P(s_k/p_i) = \frac{exp(-d(p_i, s_k))}{\sum_{j=1}^{k} exp(-d(p_i, s_j))} \tag{19}$$

$$P(p_k/s_i) = \frac{exp(-d(s_i, p_k))}{\sum_{j=1}^{k} exp(-d(s_i, p_j))} \tag{20}$$

where $P(s_k/p_i)$ represents the probability of p_i and $s_k(k = 1, 2, \ldots, k)$ being consistent. Similarly, $P(p_k/s_i)$ represents the probability of s_i and $p_k(k = 1, 2, \ldots, k)$ being consistent. When $k = i$, $P(s_i/p_i)$ and $P(p_i/s_i)$ represent the probability of predicting the same label for p_i and s_i in the corresponding cross-domain image. A higher probability indicates better recognition performance of the model for that image. By summing the negative logarithm probabilities $P(s_i/p_i)$ for all-optical images in Q_t, we obtain the alignment embedding loss for the optical images domain.

$$L_{ps} = \frac{1}{k} \sum_{i=1}^{k} -log(P(s_i/p_i)) \tag{21}$$

Similarly, the alignment embedding loss for the sketch images domain is obtained by summing the negative log probabilities $P(s_i/p_i)$ overall sketch images in Q_t:

$$L_{sp} = \frac{1}{k} \sum_{i=1}^{k} -log(P(p_i/s_i)) \tag{22}$$

Finally, we obtain the domain alignment embedding loss:

$$L = L_{ps} + L_{sp} \tag{23}$$

Experiments

We introduce a novel sketch face recognition approach that leverages Cross Grouping Transformer, aiming to rectify the limited generalization capabilities of Transformers within the context of sketch face recognition where sample sizes are constrained. In the experiments, we evaluate the performance of our method on the normal sketches dataset, the forensics dataset, and the small-scale dataset.

Dataset and setting

To verify the effectiveness of the proposed sketch face recognition algorithm RSCGT, we conduct comparative experiments on three datasets: the CUFSF dataset [25], the UoM-SGFS dataset [26], and the small-scale PRIP-VSGC dataset [26].

CUFSF dataset The CUFSF dataset is a publicly accessible collection of hand-drawn sketch face images paired with optical images. It consists of optical images from 1194 subjects sourced from the FERET database, each paired with a hand-drawn sketch image. Compared to the synthetic sketch images in the UoM-SGFS dataset, the hand-drawn sketch images in the CUFSF dataset are more similar to real optical images, with more accurate face structure and richer texture information. We randomly selected 500 pairs of optical and sketch images as the training datasets, and the remaining 694 sample pairs as testing datasets as S1.

Fig. 3 Sample of CUFSF dataset

UoM-SGFS dataset The UoM-SGFS dataset obtains criminal investigation images and the scarcity of related image information under real-world conditions. All sketches in the datasets commonly used by law enforcement agencies, and were produced under the supervision of law enforcement professionals to simulate the face sketches that might be generated in actual law enforcement scenarios. It consists of two sets: UoM-SGFS Set A and UoM-SGFS Set B. Each set contains 600 pairs of optical-sketch images. The optical images are extracted from 600 subjects in the Color FERET database, while the sketch images are synthetic and colored. We created two datasets, S2 and S3, for Set A and Set B respectively. In S2, we randomly selected 450 pairs of optical images and sketch images from UoM-SGFS Set A as the training set. The test set is composed of both the probe set and the gallery set. The probe set consists of sketch images from the remaining 150 pairs of samples in the UoM-SGFS Set A, while the gallery set includes an equal number of paired optical images, as well as an additional 1521 optical images. The additional optical images comprise 509 images from the MEDS-II database, 199 images from the FEI database, and 813 images from the LFW database. S3 was set up in the same way as S2.

PRIP-VSGC dataset The PRIP-VSGC dataset is a relatively small collection of sketch face images, consisting of 123 pairs of optical-sketch images. The optical images are sourced from 123 subjects in the AR dataset, while the sketch images are synthetic and generated using IdentityKit. The limited number of samples in the PRIP-VSGC dataset poses chal-

Table 1 Comparison results of different methods on S1

Type	Methods	Rank-1
Traditional manual-based methods	SIFT [7]	41.84
	HOG	46.03
Intra-modal methods	Fast-RSLCR [27]	75.94
	Wan's [28]	70.00
Inter-modal methods	CDD [29]	69.28
	PDT [30]	71.08
	CMTDML [31]	83.86
	CTMAN [24]	90.06
	LWVT-ResNet18 [32]	92.95
	RSCGT (Proposed method)	**93.75**

Bold highlights the most effective models or methods

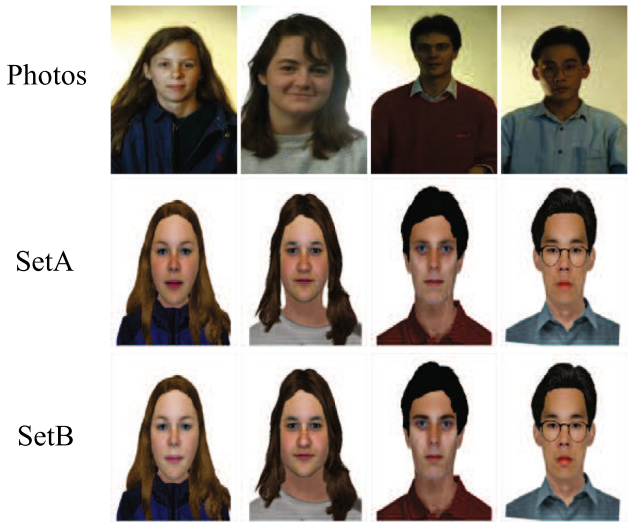

Fig. 4 Sample of UoM-SGFS dataset

Fig. 5 Sample of PRIP-VSGC dataset

lenges in terms of supporting model training and recognition tasks. For dataset S4, we randomly selected 48 sample pairs for training and the remaining 75 sample pairs for testing.

In our experiments, all datasets were pre-processed using MTCNN [33] for face detection and alignment. This pre-processing step ensured that key facial information necessary for recognition was preserved. Additionally, various data augmentation techniques, such as image warping, padding, random cropping, and horizontal flipping, were applied during both the training and testing stages. These techniques helped augment the dataset, increase its diversity, and improve the robustness and generalization of the trained models.

Experiment details

RSCGT is implemented using the deep learning PyTorch library, and the experiments require 10.9 GB of GPU memory. Several experiments were conducted to determine the hyperparameters for each component of the model. These include a batch sample size of 80 for each meta-learning task and $h = 8$ of headcount for multi-headed self-attention.

For the parametric training of the RSCGT, the CNN is initialized with the parameters of the first four stages of the network of ResNet18 pre-trained on ImageNet. The two-layer Cross Grouping Transformer and domain adaptation modules are then trained from scratch. The model parameters are iteratively updated using the AdamW optimizer [34], with an initial learning rate set to 0.00015, a momentum of 0.1, a step size of 60, and optimizer parameters $(\beta_1, \beta_2, weight\_decay) = (0.9, 0.999, 0.02)$. RSCGT is trained for 80 epochs on datasets S1 and S4, while trained for 100 epochs on S2 and S3. Each epoch consists of 100 meta-learning tasks. To ensure the reliability of the results and account for experimental fluctuations, we evaluated the model performance using a five-fold cross-validation method.

Comparisons with SOTA methods

To validate the superiority of the proposed method, we compared our approach with other sketch face recognition methods on the CUFSF dataset, the UoM-SGFS dataset, and the small-scale PRIP-VSGC dataset.

Table 2 Comparison results of different methods on S2

Methods	Rank-1	Rank-10	Rank-50
DANN [35]	52.00	85.47	95.20
CDAN [36]	57.07	88.53	95.07
BSP+CDAN [37]	55.73	89.20	97.33
SP-Net [16]	45.20	79.60	91.47
DAEN [17]	68.53	92.40	97.47
LWVT-ResNet18 [32]	73.20	93.87	98.53
RSCGT (Proposed method)	**77.60**	**94.53**	**99.07**

Bold highlights the most effective models or methods

Table 3 Comparison results of different methods on S3

Methods	Rank-1	Rank-10	Rank-50
DANN [35]	65.20	94.00	98.53
CDAN [36]	62.00	91.87	97.47
BSP+CDAN [37]	67.60	91.47	97.73
SP-Net [16]	50.93	83.07	93.20
DAEN [17]	74.00	95.20	**99.07**
LWVT-ResNet18 [32]	81.20	96.80	99.06
RSCGT (Proposed method)	**86.00**	**97.33**	98.93

Bold highlights the most effective models or methods

On the CUFSF dataset, we compared RSCGT against several other algorithms including SIFT, HOG, Fast-RSLCR [27], Wan's [28], Cross Domain Descriptor (CDD) [29], Prepended Domain Transformer (PDT) [30], Cross-modality multi-task deep metric learning (CMTDML) [31], Cascaded transformation generation network (CTMAN) [24], and LWVT-ResNet18 [32] as well as different sketch face recognition algorithms. These models perform well on the CUFSF dataset, so we selected the most discriminative Rank-1 for comparison. The results are presented in Table 1. SIFT and HOG are traditional manual-based approaches that are not very effective in extracting fixed and low-level features for recognition. Fast-RSLCR and Wan's are intra-modal algorithms that perform poorly in synthetic image recognition due to the challenges in articulating the mapping relationships between cross-modal images and synthesizing high-quality images.

CDD, CMTDML, CTMAN, and LWVT-ResNet18 are inter-modal methods that perform well in extracting modality-invariant features from cross-modal images. However, traditional CNN-based methods often struggle to capture long-range dependencies, which can limit their effectiveness in certain tasks. In comparison, our model enhances global inter-image connections through the Cross Grouping Transformer while retaining the CNN framework. This enables our model to effectively capture global relationships between features to extract modality-invariant features.

Table 4 Comparison results of different methods on S4

Type	Methods	Rank-10
Traditional manual methods	SSD [38]	45.30
	Attribute [39]	53.10
Deep learning methods	Transfer Learning [40]	52.00
	DAEN [17]	63.20
	LWVT-ResNet18 [32]	48.53
	RSCGT (Proposed method)	**64.27**

Bold highlights the most effective models or methods

Additionally, the loss function optimized by the domain adaptation module further reduces the influence of modal differences. In the experiments, RSCGT achieved the highest performance on the CUFSF dataset, outperforming the CNN-based methods by at least 3.69% on rank-1 accuracy. This demonstrates the effectiveness and superiority of RSCGT on the CUFSF dataset.

On the UoM-SGFS dataset, RSCGT was compared with SP-Net [16], Domain-Adversarial Neural Network (DANN) [35], Conditional Domain Adversarial Network (CDAN) [36], Balanced Similarity and Prediction Consistency plus Conditional Domain Adversarial Network (BSP+CDAN) [37], Domain Alignment Embedding Network (DAEN) [17] and LWVT-ResNet18 [32]. The comparison results of different methods on S2 and S3 are shown in Table 2 and Table 3.

Compared to S1, these models face more challenges in testing on S2 and S3. Therefore, we compared the performances in terms of Rank-1, Rank-10, and Rank-50. The RSCGT model exhibited exceptional performance in the sketch-based face recognition task. With a Rank-1 accuracy of 77.60% on the S2 dataset, it outperformed the DANN, CDAN, DSP+CDAN, SP-Net, and DAEN models, which achieved accuracies of 52.00%, 57.07%, 55.73%, 45.20%, and 68.53%. In the CNN-based sketch face recognition method, DAEN achieved the highest score. It utilizes which adopts the domain adaptive method and meta-learning training strategy to solve the small-sample problem, which can effectively learn the correlation information between the text and the image, and thus shows good accuracy in the sketch face recognition task. RSCGT showed a 9.07% and 12% increase in accuracy on the S2 and S3 datasets. It is demonstrated that using the ResNet18-like module as a feature extraction network, RSCGT can significantly enhance training performance. Compared to the LWVT-ResNet18, our model has a higher execution efficiency of the Transformer's self-attention mechanism and a lower overall model complexity, achieving a 4.4% improvement in Rank-1 accuracy. It also proves that the discriminative features extracted from the global information of the captured features by Cross Grouping Transformer are more beneficial for sketch face recognition.

Table 5 Results of ablation experiments with different grouping strategies

Grouping Strategies	S1			S2		
	Rank-1	Rank-10	Rank-50	Rank-1	Rank-10	Rank-50
(0, 0)	92.46	99.08	99.37	74.13	94.13	98.00
(16, 16)	**93.75**	**99.16**	**99.42**	**77.60**	94.53	**99.07**
(8, 32)	93.01	99.01	99.42	74.40	**94.80**	98.67
(4, 64)	93.23	98.96	99.42	75.47	94.53	98.80

Bold highlights the most effective models or methods

Table 6 Results of ablation experiments with different FNN structures

FFN Structures	S1			S2		
	Rank-1	Rank-10	Rank-50	Rank-1	Rank-10	Rank-50
$d_f = 4d$	91.56	98.83	99.24	71.60	93.60	98.93
$d_f = d$	92.75	99.14	99.38	74.27	94.00	99.07
$d_f = d/4$	**93.75**	**99.16**	**99.42**	**77.60**	**94.53**	**99.07**

Bold highlights the most effective models or methods

In the training of the small-sample dataset PRIP-VSGC, due to the large amount of training data required by Transformer, traditional methods with lower data requirements can achieve reasonable performance even with limited sample sizes. which are more suitable for solving small sample problems. Consequently, our model compares with several traditional methods, including Self Similarity Descriptor (SSD) [38] and Attribute [39], and with deep learning methods such as Transfer Learning [40], DAEN [17], and LWVT-ResNet18 [32]. Considering the complexity and challenges of training small-sample datasets, we choose rank-10 as the benchmark for training results. Rank-10 better reflects the performance of training results in specific contexts or scenarios, thereby avoiding the one-sidedness of over-focusing on the highest accuracy or ignoring small differences.

The experimental results are shown in Table 4. From the comparison results, the RSCGT can be better at extracting modality invariant features and enhancing the recognition rate. RSCGT outperformed other techniques in the Rank-10 recognition rate in S4, improving by at least 1.07% and up to 18.97%, demonstrating the superiority of our model. In addition, due to the limitations of Transformers in small sample training, the LVWT-ResNet18 model, which adopts the Transformer framework, does not perform well on the S4. RSCGT achieved the best performance on the small-scale PRIP-VSGC dataset with only 48 pairs of samples, demonstrating its adaptability on small datasets.

Ablation study

We perform the ablation study on the key factors in our method:

(1) *Effectiveness of grouping strategies:* To test the efficacy of grouping self-attention and cross-group self-attention in Cross Grouping Transformer, we first use the self-attention mechanism in Transformer for tests, followed by Cross Grouping Transformer. Furthermore, for experimentation with alternative grouping algorithms, we set varying numbers of groups x and tokens y inside groups.
(2) *Impact of hyper-parameters:* To verify the adaptability of the improved FFN network to sketch face recognition for small sample scenes, we adapted the intermediate layer dimension d_f of FFN to different dimensions for experiments, including the bottleneck structure of Transformer ($d_f = 4d$), the flat structure of Visual Transformer ($d_f = d$) and the improved FFN structure ($d_f = d/4$).

The results of the ablation experiment (1) are shown in Table 5. We discover that the Cross Grouping Transformer achieves better recognition results on both datasets S1 and S2 when using different grouping strategies than when using the Transformer's self-attentive mechanism $(x, y) = (0, 0)$. This result shows the efficiency of the Cross Grouping Transformer. In particular, RSCGT gets the best recognition performance when uniformly grouping $(x, y) = (16, 16)$, and the recognition accuracy of Rank-1 is enhanced by 1.29% on S1 and 3.47% on S2 when compared to Transformer. This result proves the superiority of Cross Grouping Transformer.

The results of the ablation experiment (2) are shown in Table 6. We found that the model achieved better recognition results on datasets S1 and S2 when the dimension of the intermediate layer of FNN was decreased. We also discover that compared with $d_f = 4d$ in Transformer and the commonly used $d_f = d$, the improved FFN ($d_f = d/4$) achieves the best results. Following the experiments, the improved FFN enhanced the Rank-1 recognition accuracy by 2.19% and 1%

for dataset S1 and 6% and 3.33% for dataset S2, respectively. These experimental results demonstrate the effectiveness of the improved FFN and show that the improved FFN structure is more adaptable to the small sample scenario of sketch face recognition.

Conclusion

In this paper, we propose a novel Residual Serialized Cross Grouping Transformer (RSCGT) architecture to address the issues of excessive complexity and modality differences in Transformer models for sketch face recognition. A residual serialized module is proposed to reduce the computational cost of Transformer models by efficiently embedding high-dimensional image features. A two-layer Cross Grouping Transformer module is proposed to capture contextual relationships of features. We adopt a domain adaptive module to handle the modality differences between sketch images and optical images. Additionally, we adopt an effective meta-learning training strategy specifically designed for small samples to prevent overfitting issues.

Acknowledgements This work was supported by the National Natural Science Foundation of China (No.62201066, No.62001033, U20A20163).

Data Availability The datasets analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Conflict of interest The authors have no Conflict of interest in the publication of this paper.

References

1. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. Adv Neural Inform Process Syst 30
2. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, et al (2020) An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929
3. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B (2021) Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022
4. Srinivas A, Lin T-Y, Parmar N, Shlens J, Abbeel P, Vaswani A (2021) Bottleneck transformers for visual recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16519–16529
5. Wang H, Zhu Y, Green B, Adam H, Yuille A, Chen L-C (2020) Axial-deeplab: Stand-alone axial-attention for panoptic segmentation. In: European Conference on Computer Vision, pp. 108–126. Springer
6. Klare B, Li Z, Jain AK (2010) Matching forensic sketches to mug shot photos. IEEE Trans Pattern Anal Mach Intell 33(3):639–646
7. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vis 60(2):91–110
8. Ojala T, Pietikainen M, Maenpaa T (2002) Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans Pattern Anal Mach Intell 24(7):971–987
9. Han H, Klare BF, Bonnen K, Jain AK (2012) Matching composite sketches to face photos: a component-based approach. IEEE Trans Inform Foren Secur 8(1):191–204
10. Bonnen K, Klare BF, Jain AK (2012) Component-based representation in automated face recognition. IEEE Trans Inform Foren Secur 8(1):239–253
11. Gower JC (1975) Generalized procrustes analysis. Psychometrika 40:33–51
12. Zhang L, Lin L, Wu X, Ding S, Zhang L (2015) End-to-end photo-sketch generation via fully convolutional representation learning. In: Proceedings of the 5th ACM on International Conference on Multimedia Retrieval, pp. 627–634
13. Mahfoud S, Daamouche A, Bengherabi M, Hadid A (2022) Hand-drawn face sketch recognition using rank-level fusion of image quality assessment metrics. Bull Polish Acad Sci Tech Sci 70(6)
14. Wan W, Gao Y, Lee HJ (2019) Transfer deep feature learning for face sketch recognition. Neural Comput Appl 31:9175–9184
15. Gui, S., Wang Z, Chen J, Zhou X, Zhang C, Cao Y (2023) Mt4mtl-kd: a multi-teacher knowledge distillation framework for triplet recognition. IEEE Trans Med Imaging
16. Cheraghi H, Lee HJ (2019) Sp-net: a novel framework to identify composite sketch. IEEE Access 7:131749–131757
17. Guo Y, Cao L, Chen C, Du K, Fu C (2020) Domain alignment embedding network for sketch face recognition. IEEE Access 9:872–882
18. Chakravarthi B, Ng S-C, Ezilarasan M, Leung M-F (2022) Eeg-based emotion recognition using hybrid cnn and lstm classification. Front Comput Neurosci 16:1019776
19. Chen H, Wang Y, Guo T, Xu C, Deng Y, Liu Z, Ma S, Xu C, Xu C, Gao W (2021) Pre-trained image processing transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12299–12310
20. Ying C, Cai T, Luo S, Zheng S, Ke G, He D, Shen Y, Liu T-Y (2021) Do transformers really perform badly for graph representation? Adv Neural Inform Process Syst 34:28877–28888
21. Wu B, Xu C, Dai X, Wan A, Zhang P, Yan Z, Tomizuka M, Gonzalez JE, Keutzer K, Vajda P (2021) Visual transformers: where do transformers really belong in vision models? In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 599–609
22. Mehta S, Koncel-Kedziorski R, Rastegari M, Hajishirzi H (2019) Define: Deep factorized input token embeddings for neural sequence modeling. arXiv preprint arXiv:1911.12385
23. Sangkloy P, Lu J, Fang C, Yu F, Hays J (2017) Scribbler: controlling deep image synthesis with sketch and color:5400–5409

24. Cao L, Huo X, Guo Y, Du K (2021) Sketch face recognition via cascaded transformation generation network. IEICE Trans Fund Electron Commun Comput Sci 104(10):1403–1415
25. Zhang W, Wang X, Tang X (2011) Coupled information-theoretic encoding for face photo-sketch recognition. In: CVPR 2011, pp. 513–520. IEEE
26. Galea C, Farrugia RA (2016) A large-scale software-generated face composite sketch database. In: 2016 International Conference of the Biometrics Special Interest Group (BIOSIG), pp. 1–5. IEEE
27. Wang N, Gao X, Li J (2018) Random sampling for fast face sketch synthesis. Elsevier
28. Wan W, Lee HJ (2019) Generative adversarial multi-task learning for face sketch synthesis and recognition. In: 2019 IEEE International Conference on Image Processing (ICIP), pp. 4065–4069. IEEE
29. Kumar VA, Rajesh K, Antony R (2021) Cross domain descriptor for face sketch-photo image recognition. In: 2021 2nd International Conference on Advances in Computing, Communication, Embedded and Secure Systems (ACCESS), pp. 228–231. IEEE
30. George A, Mohammadi A, Marcel S (2022) Prepended domain transformer: heterogeneous face recognition without bells and whistles. IEEE Trans Inform Foren Secur 18:133–146
31. Feng Y, Wu F, Huang Q, Jing X-Y, Ji Y, Yu J, Chen F, Han L (2019) Cross-modality multi-task deep metric learning for sketch face recognition. In: 2019 Chinese Automation Congress (CAC), pp. 2277–2281. IEEE
32. Cao L, Yin J, Guo Y, Du K, Zhang F (2023) Sketch face recognition based on light semantic transformer network. IET Compute Vis 17(8):962–976
33. Xiang J, Zhu G (2017) Joint face detection and facial expression recognition with mtcnn. In: 2017 4th International Conference on Information Science and Control Engineering (ICISCE), pp. 424–427. IEEE
34. Loshchilov I, Hutter F (2017) Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101
35. Ganin Y, Ustinova E, Ajakan H, Germain P, Larochelle H, Laviolette F, March M, Lempitsky V (2016) Domain-adversarial training of neural networks. J Mach Learn Res 17(59):1–35
36. Long M, Cao Z, Wang J, Jordan MI (2018) Conditional adversarial domain adaptation. Adv Neural Inform Process Syst 31
37. Chen X, Wang S, Long M, Wang J (2019) Transferability vs. discriminability: Batch spectral penalization for adversarial domain adaptation. In: International Conference on Machine Learning, pp. 1081–1090. PMLR
38. Mittal P, Jain A, Goswami G, Singh R, Vatsa M (2014) Recognizing composite sketches with digital face images via ssd dictionary. In: IEEE International Joint Conference on Biometrics, pp. 1–6. IEEE
39. Mittal P, Jain A, Goswami G, Vatsa M, Singh R (2017) Composite sketch recognition using saliency and attribute feedback. Inform Fus 33:86–99
40. Mittal P, Vatsa M, Singh R (2015) Composite sketch recognition via deep network-a transfer learning approach. In: 2015 International Conference on Biometrics (ICB), pp. 251–256. IEEE