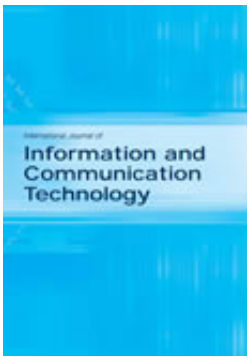
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642
https://www.inderscience.com/ijict

A light end-to-end comprehensive attention architecture for advanced face parsing

Cong Han, Peng Cheng, Zhisheng You

Article History:

Received:	29 October 2024
Last revised:	19 December 2024
Accepted:	19 December 2024
Published online:	13 February 2025

A light end-to-end comprehensive attention architecture for advanced face parsing

Cong Han

College of Computer Science,
Sichuan University,
Chengdu, Sichuan Province, 610065, China
Email: colinhancong@stu.scu.edu.cn

Peng Cheng*

School of Aeronautics and Astronautics,
Sichuan University,
Chengdu, Sichuan Province, 610065, China
Email: chengpeng@scu.edu.cn
*Corresponding author

Zhisheng You

College of Computer Science,
Sichuan University,
Chengdu, Sichuan Province, 610065, China
Email: youzhisheng@scu.edu.cn

Abstract: Face parsing involves segmenting a face into various semantic regions, but challenges such as complex structures, varying poses, and occlusions make achieving high performance difficult, even for state-of-the-art methods. To address these challenges, we propose FP-Transformer, a novel architecture that combines CNNs and Transformer to extract both long-range and short-range features. Our design includes: 1) a U-shaped Encoder-Decoder with hierarchical feature fusion and hybrid attention blocks for semantic learning; 2) convolution-based patch embedding and merging to retain edge information; 3) a novel Bunch-layer normalisation (BLN) to maintain consistent normalisation across patches. Experiments on CelebAMask-HQ and LaPa datasets demonstrate the effectiveness of our approach, achieving mean F1 scores of 87.1% and 92.6%, respectively. Our model performs robustly even under occlusions, extreme poses, and complex backgrounds.

Keywords: face parsing; face analysation; face segmentation; self-attention mechanism.

Reference to this paper should be made as follows: Han, C., Cheng, P. and You, Z. (2025) 'A light end-to-end comprehensive attention architecture for advanced face parsing', *Int. J. Information and Communication Technology*, Vol. 26, No. 3, pp.89–109.

Biographical notes: Cong Han is a PhD candidate in College of Computer Science, Sichuan University, Chengdu, China. In 2018, he received a dual Bachelor's degree in Computer Science and Technology and Financial Engineering from the same institution. His research interests include computer vision and face recognition.

Peng Cheng is an Associate Professor with College of Aeronautics and Astronautics in Sichuan University, Chengdu, China, received Ph.D. degree from Sichuan University. His research interests include computer vision, image fusion and image processing.

Zhisheng You is a Professor with College of Computer Science, Sichuan University, Chengdu, China. He is Deputy Director of the Academic Committee of National Key Laboratory of Fundamental Science on Synthetic Vision, Sichuan University. His research interests include computer applications, graphics and image processing, air traffic control technology, visual synthesis and real-time software engineering.

1 Introduction

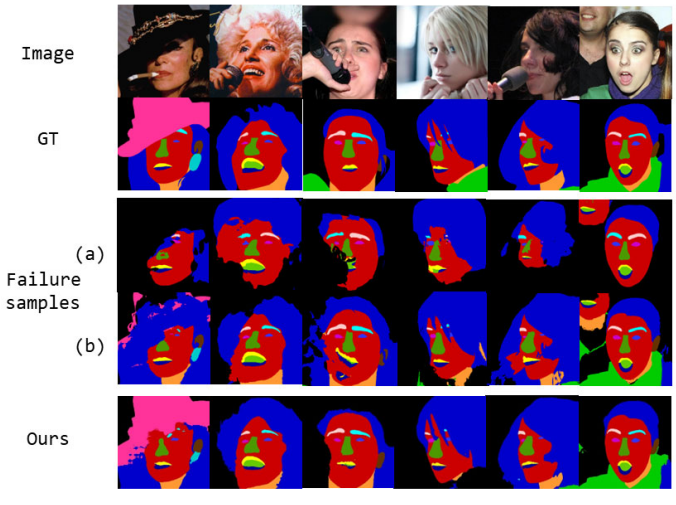
Face parsing, which refers to the task of segmenting a face into different semantic regions, has become increasingly important in various computer vision applications. It is a fundamental task that provides the basis for a wide range of downstream applications, such as facial analysis (Zheng et al., 2021; Wood et al., 2021), face recognition (Umirzakova and Whangbo, 2022), and face editing (Lee et al., 2020). A good face parsing method can accurately partition a face into several regions, including the skin, hair, eyes, nose, mouth, and eyebrows, and can provide rich information about the structure and appearance of a face.

With the rapid advancement of deep learning, deep convolutional neural networks (CNNs) have become a popular choice due to their excellent feature learning capabilities. This has led to their widespread use in various facial recognition tasks (Jayaraman et al., 2020), especially in face parsing task (Guo et al., 2018; Jackson et al., 2016; Lin et al., 2021; Liu et al., 2017, 2015; Luo et al., 2020; Zhou et al., 2017, 2015). Despite recent advances in face parsing CNN-based methods, there are still some challenges that need to be addressed. One of the main challenges is small receptive field of convolution operators, make the methods (Jackson et al., 2016; Lin et al., 2021; Liu et al., 2017; Luo et al., 2020; Zhou et al., 2017; Guo et al., 2018), based on CNN lack the ability to model long-range contextual information, which can hinder face segmentation performance. For instance, existing methods fail to accurately parse faces that are captured at extreme angles or that are partially occluded by objects or other people, as shown in Figure 1 row (a). Row (a) contains the failed samples of Y. Lin's method (Lin et al., 2021). To address this issue, Jackson et al. (2016), Liu et al. (2015) and Zhou et al. (2017) combine CNNs and CRFs to learning the long-range information. However, these methods do not consider the correlation among various objects. Te et al. (2020), Zheng et al. (2022) and Te et al. (2021) exploit the relations between regions for face parsing by modelling GCN (Zhang et al., 2019), which shows a great performance on reasoning the region-level information to get the information between different facial parts. Nevertheless, to enhance performance, the sizes of these models and their computational complexity have

increased significantly. The numbers of parameters and MACs for EHANet (Luo et al., 2020), which is based on CNN, are 43.9M and 11.5G calculated by using pytorch-opcounter. The numbers of parameters for EAGR (Te et al., 2020), AGRNet (Zheng et al., 2022), DML-CSR (Te et al., 2021) with GCN are 66.61 M, 66.42 M and 67.94 M. MACs of them are 236.41 G, 205.14 G and 252.58 G. However, these methods still predict some failure sample, as shown in Figure 1 row (b). Row (b) contains the failed samples of DML-CSR (Te et al., 2021), shows that in scenarios with complex backgrounds, low lighting and facial obstructions, existing methods often become misled, confusing actual facial features with similar elements in the background or obstructions in front of the face.

To address these issues, we design a transformer-based model to improve long-range information extraction ability of face parsing model. Transformer-based models, are a popular type of deep learning models that have been proposed for natural language processing tasks (Wu, 2024). Inspired by the Transformer architecture, the development of the vision transformer (ViT) architecture (Dosovitskiy et al., 2020), has shown promising results in various computer vision tasks by dividing images into non-overlapping patches and treats them as sequence elements, including image classification (Wu et al., 2021; Liu et al., 2021; Wang et al., 2021; Touvron et al., 2021; Chen et al., 2021), object detection (Dai et al., 2021a, 2021c; Zhu et al., 2020), face recognition (Ge et al., 2023; Song et al., 2022) and semantic segmentation (Zheng et al., 2020; Xie et al., 2021; Woo et al., 2023; Cheng et al., 2022; Mallick et al., 2023; Zhang et al., 2023). However, few have applied transformer-based models to facial parsing tasks. FaRL-B (Zheng et al., 2021) relies on huge Transformer model and both image and language datasets for multimodal training, which is hard to train and inference. Consequently, there is a need for a novel architecture tailored specifically for face parsing tasks that addresses these limitations while harnessing the power of the Transformer architecture.

Figure 1 Failure samples from existing face parsing methods (see online version for colours)



In this paper, we propose a novel end-to-end face parsing method based with Multi-Head-Self-Attention mechanism (Wu, 2024). Our model is a U-shaped Encoder-Decoder architecture (Siddique et al., 2021) that incorporates hierarchical feature extraction. The Multi-Head Self-Attention mechanism (Wu, 2024) and CNN are combined in each attention block to capture both long-range and short-range relationships in face images. Additionally, the patch merging operation (Liu et al., 2021; Wang et al., 2021) and non-overlapping patch embedding (Dosovitskiy et al., 2020), which are ineffective at capturing positional information for different patches and fine-grained details, are replaced by CNN to retain the position information between different image patches or tokens. To decrease the parameters of our model, we adopt the window self-attention. Meanwhile, due to the integration of CNNs, fewer attention blocks are needed for shifting attention windows, as in Swin (Liu et al., 2021), which significantly decreases the complexity of our model. Additionally, during our experiment, we observed that the original LayerNorm employed in the ViT architecture (Dosovitskiy et al., 2020) normalises the embeddings of individual image patches independently. Therefore, we propose a new normalisation layer, termed bunch layer normalisation (BLN), designed to replace LayerNorm and optimise internal covariate shift and thereby ensure the normalisation consistency of the relevant patches (tokens) and improve the model's overall performance. To further enhance the model's representation capacity, we employ feature fusion module (FFM) (Dai et al., 2021b) that merge features across different resolutions. Our proposed innovative aggregation of techniques aims to address the small receptive field limitation associated with CNNs, as well as the computational intensity and complexity commonly found in ViT architectures, making our model both efficient and effective for face parsing tasks.

The main contributions of this paper are as follows:

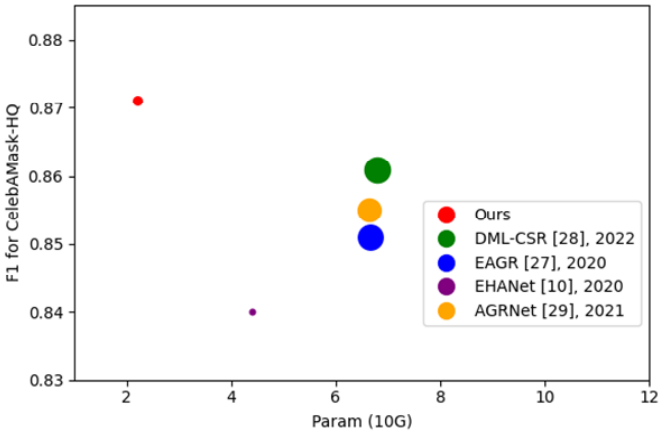
1 We introduce a hybrid, end-to-end face parsing model that fuses CNNs and Transformers, influencing the attention mechanism for short-range feature capture, and augment the novel Long-short block with U-shape and convolution layers for fine-grained feature extraction.

2 Convolutional patch merging operation and convolutional patch embedding are designed to retain the positional information between different image patches or tokens, which are crucial for precise face parsing but are often lost in the original patch merging and patch embedding processes.

3 BLN has been designed to replace LayerNorm. BLN executes normalisation across all correlated patches, which is an operation more frequently encountered in computer vision tasks.

4 Our method attains an average F1 score of 92.6% on the LaPa dataset and 87.1% on the CelebAMask-HQ dataset, all while utilising fewer computational resources, shown as Figure 2. The circle size of our model is smaller than others, which means our model has less MACs counts. A wide range of experiments confirms the efficiency and robustness across a range of face parsing tasks of our proposed method.

Figure 2 Model performance and complexity (see online version for colours)



2 Related work

Face Parsing: The field of image segmentation (Zhao et al., 2023; Ji and Zhong, 2024; Minaee et al., 2020; Yu et al., 2023) has seen considerable advancements in recent years, with deep neural network being proposed. The existing literature on face parsing is extensive and focuses particularly on utilising CNN based model. Zhou et al. (2017) incorporated features derived from CNN into the conditional random field (CRF) framework to characterise individual pixel labels and their adjacent relationships. Jackson et al. (2016) introduces a CNN cascade that uses pose-specific landmarks for semantic part segmentation, marking the first exploration of the relationship between pose estimation and segmentation. Luo et al. (2012) constructed a hierarchical framework by integrating multiple independent deep networks. Zhou et al. (2015) employed a concatenation of input image pyramids with various feature maps to enhance scale invariance within the network. They introduced an interlinked convolution neural network (iCNN) for face parsing. The iCNN consists of multiple convolution layer taking input in different scales, and a special interlinking layer allows the CNN to exchange information, enabling them to integrate local and contextual information efficiently. The model uses extensive downsampling and upsampling in the interlinking layers, which is different from traditional CNN. Lin et al. (2019) proposed a novel RoI Tanh-warping operator that combines central and peripheral vision. This operator addresses the challenge of focusing on a limited region of interest (RoI) while also considering an unpredictable surrounding context. Their hybrid convolution neural network (CNN) for face parsing uses local methods for inner facial components and global methods for outer facial components, providing a balanced approach to face parsing. Nevertheless, CNN architectures exhibit limitations in capturing global or long-term feature representations, primarily due to constraints associated with their receptive field. In response to this challenge, several researchers have proposed alternative architectural designs to augment the capability of long-term feature extraction. Liu et al. (2017) introduced a face parsing algorithm that combines hierarchical representations learned by a CNN and accurate label propagation achieved by a recurrent neural network (RNN). Their RNN-based

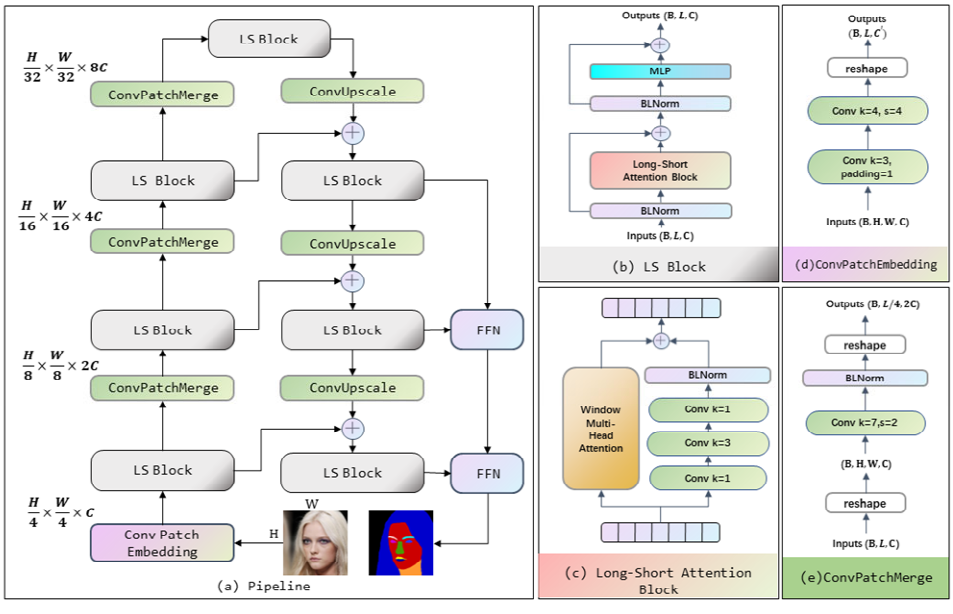
propagation approach enables efficient inference over a global space with the guidance of semantic edges generated by a local convolution model, offering a unique perspective on face parsing. Te et al. (2020) import graph convolution network (GCN) into face parsing task. They proposed adaptive graph representation learning and graph reasoning over facial components. Edge Aware Graph Reasoning (EAGR) module, proposed in their paper, learns representative vertices that describe each component and exploits the component-wise relationship to produce accurate parsing results from face image and edge map. Zheng et al. (2022) introduces the decoupled multi-task learning with cyclical self-regulation (DML-CSR) for face parsing, addressing challenges like spatial inconsistency and boundary confusion by employing a multi-task model and a dynamic dual GCN. The method utilises cyclical self-regulation for model refinement and has demonstrated state-of-the-art performance on datasets such as CelebAMask-HQ, and LaPa. These studies highlight the significance of establishing connections between various facial components, a fundamental aspect of our approach. However, for face images with extreme poses, occlusions, and complex backgrounds, these methods still incorrectly segment faces, while also experiencing increased computational complexity and model sizes.

Vision transformer: Dosovitskiy et al. (2020) is the first to employ a pure transformer structure, termed vision transformer (ViT), directly to feed sequences of image patches for computer vision tasks in ViT. Specifically, ViT breaks down each image into a fixed-length sequence of tokens (non-overlapping patches) and then employs multiple standard Transformer layers. These layers include the multi-head self-attention module (MHSA) and the positionwise Feed-forward module (FFN) to process these tokens. Demonstrating superior performance compared to existing state-of-the-art convolution networks, ViT achieves these results with significantly less training resources, especially when pre-trained on extensive datasets and adapted to various image recognition benchmarks. In order to reduce computational complexity, researchers have attempted to modify ViT by utilising experience from the CNN structure. Pyramid vision transformer (PVT), is a backbone network for various dense prediction tasks, designed by Wang et al. (2021) to overcome the limitations of using Transformers like ViT for vision tasks. PVT is capable of high-resolution outputs and reduced computational costs by using a progressive shrinking pyramid. Yuan et al. (2021) finds the origin tokenisation of input images fails to model the important local structure. Motivated by CNN architecture, they propose tokens-to-token ViT, which can structure the image to tokens by recursively aggregating surround tokens into one token. Touvron et al. (2021) delves deeper into data-efficient training and distillation for ViT. This study explores the efficient integration of CNNs and Transformers to effectively model both local and global dependencies for image classification. The conditional position encoding visual transformer (CPVT), designed by Chu et al. (2021), substitutes the predefined positional embedding in ViT with conditional position encoding (CPE). This adaptation allows Transformers to handle input images of any size without the need for interpolation. Wu et al. (2021) introduce convolutional token embedding and convolutional projection in ViT to employs all the benefits of CNNs: local receptive field, shared weights and spatial subsampling, while keeping advantages of Transformer. Liu et al. (2021) presented the Swin-Transformer, incorporating a window multi-head self-attention (W-MSA) module with relative position bias, aimed at mitigating computational complexity. Furthermore, they indicated that with appropriate adjustments, W-MSA can achieve performance equal to or even surpassing that of traditional global self-attention after reducing computational

complexity. FaRL-B is one of the few successful cases of applying transformers to face parsing, thanks to the highly parameterised ViT and extensive multi-modal training with both image and language datasets. Apparently, the size of FaRL-B and the vast amount of training data make this approach highly impractical to implement.

3 Methodology

Figure 3 The network architecture of the components of FP-transformer (see online version for colours)



3.1 Overall

FP-Transformer is inspired by the popular Swin-Transformer (Liu et al., 2021), which pioneered the concept of shifted window attention and UniNext's (Lin et al., 2023) validation of window attention further confirms its effectiveness, supporting the approach used in our face parsing method. Meanwhile several modified improvements have been incorporated in our proposed architecture. Figure 3(a) shows the overall pipeline of FP-transformer. It is a modified U-shaped network with our LS block, which contains a long-short attention block. We just give a face image I(x) and face mask groudtruth M(x), and calculate the Focal loss (Lin et al., 2016) to train our model. Convolutional patch embedding, convolutional patch merging, long-short attention block, and BLN will be introduced in the following sections.

3.2 Convolution operation

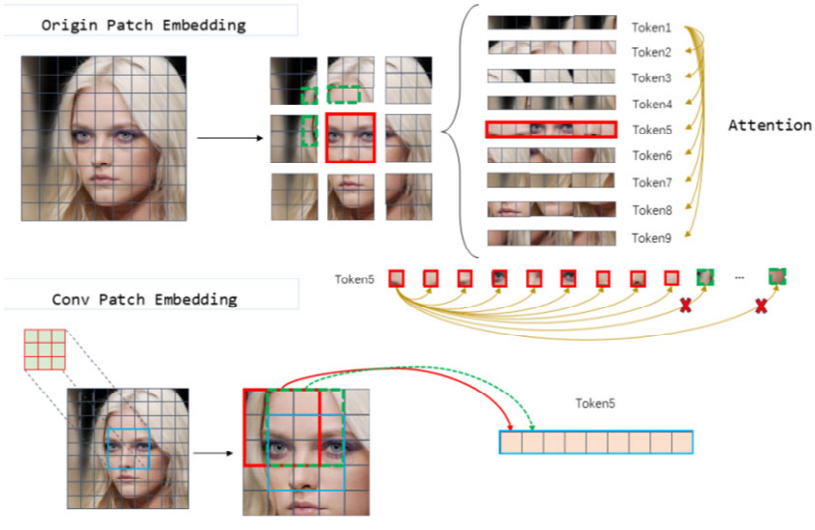
Contrary to the initial patch embedding and patch merging used in the Swin Transformer, our approach employs convolution layers to strengthen the connectivity among adjacent pixels, which tends to be disrupted when the image is directly divided into smaller patches. As shown in Figure 4, for origin patch embedding method, only the relationships between pixels within a patch are preserved, while the information of adjacent pixels outside of the patch (indicated by the red and green parts) is discarded. The lost information is crucial for face segmentation, which depends on the connectivity between pixels, and its absence adversely affects segmentation accuracy. Consequently, desired by Wu et al. (2021), we replace the original patch embedding with a convolution layer capable of gathering information from surrounding pixels. Traditional patch embedding simply rearranges the data, while convolutional patch embedding introduces additional computational overhead. However, for image-related tasks, convolutional patch embedding proves more effective as it better preserves the relational information between image patches. The convolution layer is brought in not only to improve the extraction of local spatial features by combining a pixel's information with its surroundings but also allows us to abandon position embedding, due to the structural information preserved by the convolution operation. Formally, given a face image $X_{face} \in \mathbb{R}^{H\times W\times C}$ as input, $f(\ldots)$ is convolution layer with $s \times s$ kernel, $\left\lceil \frac{s}{2} \right\rceil$ stride and $\left\lfloor \frac{s}{2} \right\rfloor$ padding. The token map is $f(X_{Image}) \in \mathbb{R}^{\hat{H}\times\hat{W}\times\hat{C}}$ with height and weight

$$\hat{H} = \left\lfloor \frac{H + 2\left\lfloor \frac{2}{s} \right\rfloor - s}{\left\lceil \frac{s}{2} \right\rceil} + 1 \right\rfloor, \hat{W} = \left\lfloor \frac{W + 2\left\lfloor \frac{2}{s} \right\rfloor - s}{\left\lceil \frac{s}{2} \right\rceil} + 1 \right\rfloor \tag{1}$$

For the patch merging process, our objective remains consistent. We employ convolution layers to simultaneously extract short-range features and reduce the size of the feature map. Both convolutional patch embedding and convolutional patch merging can improve the overall model's ability to interpret the information among adjacent pixels.

When a convolutional layer is added to the patch embedding and patch merging process, each patch or token contains information from both the positionally adjacent patches or tokens and itself. There is no need to shift attention window to contain the long-range information extraction ability, as in Swin transformer. Therefore, while a single LS block is used to extract features for different resolution feature maps, the corresponding Swin architecture requires two Swin blocks. If we use original Swin-T for 448 × 448 inputs to construct a U-shaped model, the number of parameters and MACs will be about 41.3 M and 35.1 G, which is more complex than ours.

Figure 4 The difference between origin patch embedding and convolutional patch embedding (see online version for colours)



3.3 Bunch layer normalisation

Layer normalisation (LN) (Ba et al., 2016) is the common choice in ViT models. The formulation of LayerNorm used in origin ViT can be expressed as:

$$y = \frac{x - E[x]}{\sqrt{Var[x] + \varepsilon}} * \gamma + \beta, x \in \mathbb{R}^{b \times l \times c} \tag{2}$$

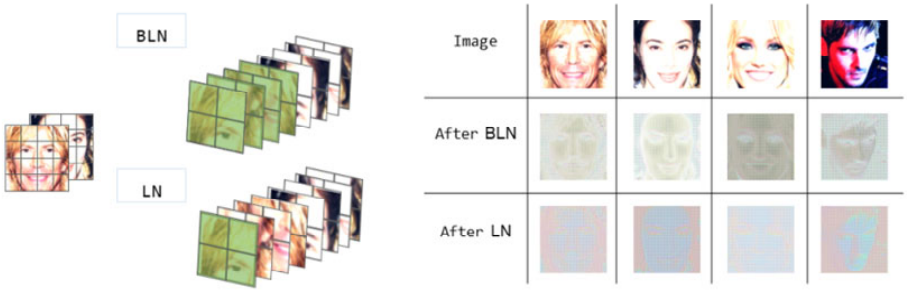
where x is the input feature map, $E[x]$ and $Var[x]$ means the mean and standard deviation computed along the axis c, γ and β are learnable parameters, and $l = h \times w$ (height and width). In a ViT, the input feature map consists of a series of tokens derived from tokenising image patches. Consequently, each token originating from the same image exhibits distinct expected $E[x]$ and $Var[x]$. This means that identical regions of a facial image, when tokenised into separate tokens, will experience varying alterations as they pass through LN. In other way, the contrast and luminance relationships among image tokens become obscured after LN., and this is the reason why we utilise BLN in our paper. In BLN, we compute mean and standard deviation on tokens from same image. The formulation can be expressed as:

$$y = \frac{x_i - E\left[\sum_{i=1}^{m} x_i\right]}{\sqrt{Var\left[\sum_{i=1}^{m} x_i\right] + \varepsilon}} * \gamma + \beta, x_i \in X, X \in \mathbb{R}^{b \times \frac{l}{m} \times c} \tag{3}$$

where x_i is a patch of an image, X is an image sample. The distinctions between BLN and LN are illustrated, and the impact of these two normalisation methods is demonstrated in Figure 5. The input image is divided into patches and separately processed through BLN and LN layers. According to Figure 5, the reconstructed outputs reveal that LN obscures

facial details, making key features like eyes, nose, and mouth difficult to discern. In contrast, although some texture information is lost with BLN, it retains the structural information of faces. Facial components have more defined edges, preserving crucial information for segmentation tasks.

Figure 5 Illustration of BLN and LN, and the contrasting effects on the feature map subsequent to each normalisation layer (see online version for colours)



This indicates BLN is more suited for computer vision tasks where local spatial relationships are important. The facial reconstructions demonstrate these differences, with BLN better retaining semantic facial structures.

3.4 Long-short range attention block

Multi-headed self-attention (MHSA) (Wu, 2024) is the core component of the attention block in the original Transformer architecture. MHSA enables the model to learn multiple representations and capture different contextual relationships for each token. By using multiple parallel attention heads, rather than a single head, MHSA allows the Transformer to encode more complex interactions and subtle nuances in the input data. This gives the transformer greater expressive power to model the myriad relationships between image tokens in an image. The flexibility of multi-headed attention provides the Transformer with much of its representational capacity and has been key to its effectiveness on many sequence modelling tasks. The formula of attention is:

$$MultiHead(Q,K,V) = Concat\left(head_1,...,head_h\right)W^O \tag{4}$$

$$Attention\left(QW_i^Q, KW_i^K, VW_i^V\right) = softmax\left(\frac{\left(QW_i^Q\right)\left(KW_i^k\right)^T}{\sqrt{d_k}}\right)VW_i^V \tag{5}$$

where $Q,K,V \in \mathbb{R}^{b\times l\times d}$ is projection of input feature map $X \in \mathbb{R}^{b\times h\times w\times c}$, d is the dimension of the feature, projections are parameter matrices $W_i^Q \in \mathbb{R}^{d_{model}\times d_k}$, $W_i^K \in \mathbb{R}^{d_{model}\times d_k}$, $W_i^V \in \mathbb{R}^{d_{model}\times d_v}$ and $W^O \in \mathbb{R}^{hd_v\times d\ model}$.

Our proposed architecture differs from the traditional transformer block. As mentioned, the attention block in transformers models relationships between tokens rather than information within each token. In addition to the ConvPatchEmbedding and ConvPatchMerge layers we introduce, we design a parallel structure of CNN and

attention blocks. This allows the model to jointly extract both inter-token relationships via attention, and feature representations within tokens via CNN. While the attention block models contextual interactions between tokens which is long-range feature, the CNN extracts localised visual features within each token which is short-range feature. By complementing attention with a CNN branch, our model can simultaneously learn cross-token relationships (long-range feature) and fine-grained local details (short-range feature). This hybrid CNN and attention design represents a departure from relying only on self-attention in standard transformer blocks.

As illustrated in Figure 3, we utilise a bottleneck structure (He et al., 2015) in the CNN branch to reduce the feature dimension. This bottleneck design, commonly employed in ResNet architectures, uses a 1 × 1 convolution to decrease the number of channels followed by 3 × 3 convolutions. The bottleneck allows the CNN branch to operate on lower-dimensional feature maps, decreasing computational cost. By referencing the bottleneck structure from ResNet, we provide clarity on how the feature dimensions are adjusted within the CNN component of our proposed model.

Let x denote the input feature map, Δ is the operation to reshape $x \in \mathbb{R}^{b\times l\times c}$ into $x \in \mathbb{R}^{b\times l\times w\times c}$, and Δ^{-1} means reshaping $x \in \mathbb{R}^{b\times h\times w\times c}$ into $x \in \mathbb{R}^{b\times l\times c}$. The forward pass of bottleneck brunch can be described as:

$$x' = GELU\left(BLN\left(\Delta\left(bottleneck\left(\Delta^{-1}(x)\right)\right)\right)\right) \tag{6}$$

and the forward pass of the long-short range attention block can be described as:

$$\begin{aligned} &Q, K, V = \Phi_q(x), \Phi_k(x), \Phi_v(x) \\ &x_{output} = MutliHeadAttention(Q, K, V) + x' + x \end{aligned} \tag{7}$$

Window attention was introduced by Liu et al. (2021) in the Swin Transformer architecture to reduce the computational complexity of global self-attention. This is done by limiting self-attention to local, non-overlapping windows of M x M patches. The authors also used shifted windows between stages to enable connectivity between non-overlapping windows. However, in our proposed model, we leverage several other techniques to capture inter-window information. Specifically, our ConvPatchEmbedding, ConvPatchMerge layers, and bottleneck structure within our attention block provide mechanisms to extract features across windows. Therefore, we can simply utilise fixed window attention without shifted windows while still encoding relationships between different spatial regions. The proposed components allow connectivity between non-overlapping windows, reducing the need for explicitly linking windows as done in Swin Transformer. Our model extracts inter-window information via convolution operations rather than shifted windows. We analyse the FLOPs of using two Swin transformer blocks versus long-short range attention block to compare the computational complexity of the two structures. Specifically, the FLOPs for two Swin transformer blocks are $24 \times H \times W \times C^2 + 4 \times M^2 \times H \times W \times C$, while the combination of one Swin transformer block and one long-short range attention block reduces the complexity to $17 \times H \times W \times C^2 + 2 \times M^2 \times H \times W \times$ C. The results show that two Swin transformer blocks have higher complexity as they involve more C^2 and M^2C terms. This indicates that introducing the long-short range attention block can effectively reduce computational complexity while potentially providing performance improvements.

3.5 Feature fusion

Because of the ConvPatchMerge operation in our hierarchical architecture, information for higher resolution feature maps can be dropped. These features provide necessary information for generating face parsing prediction at corresponding resolution. Taking inspiration from U-net (Ronneberger et al., 2015), we use skip connection between corresponding resolution feature maps to allow model to propagate lost information, shown in Figure3(a). Furthermore, during our experiments, we found that doing feature fusion among different resolution in up-scaling process can improve the model performance. We use attention feature fusion (Dai et al., 2021b) as feature fusion module using multi-scale channel attention. To more comprehensively integrate the two different features, we combine global attention and local attention to obtain enriched feature representations, better capturing complex contextual information. Based on these features, we compute the weights of the two features, enhancing the differences and complementarities between them, thereby achieving a more effective feature fusion. Specifically, given feature maps x_1 and x_2 (x_2 is the smaller one), we firstly use a convolution layer $f_{proj}(\cdot)$ to scale x_2 to the same resolution as x_1. Then the multi-scale channel attention can be presented as:

$$\mathbf{M}\left(x_1 + x_2\right) = \mathbf{L}\left(x_1 + x_2\right) \oplus \mathbf{g}\left(GAP\left(x_1 + x_2\right)\right) \tag{8}$$

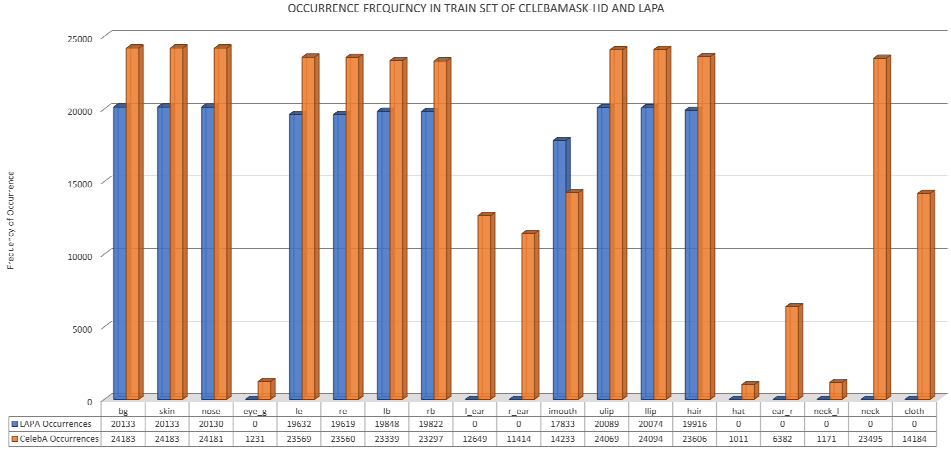
where $\mathbf{L}(\cdot)$ denotes local attention (Conv-Bn-Relu-Conv), $\mathbf{g}(\cdot)$ denotes global attention (Conv-Bn-Relu-Conv), GAP($\cdot$) is global average pooling. And the fused feature Z can be describe as:

$$Z = \mathbf{M}\left(x_1 + x_2\right) \otimes x_1 + \left(1 - \mathbf{M}\left(\left(x_1 + x_2\right)\right)\right) \otimes x_2 \tag{9}$$

4 Experiments

Datasets. Our experiments performed on CelebAMask-HQ (Lee et al., 2020) and LaPa (Liu et al., 2020). The CelebAMask-HQ is one of the most commonly used datasets for face parsing and it contains 24,183, 2,993, 2,824 images for training, validation and testing with labels: background, facial skin, left or right brow, left or right eye, nose, upper or lower lip, inner mouth, hair, left or right ear, eye-glass, earring, necklace, neck and cloth. The LaPa dataset consists of more than 22,000 facial images with abundant variations in expression, pose and occlusion, and each image of LaPa is provided with a 11-category pixel-level labels including the first ten labels of CelebAMask-HQ dataset. And there are 18,176 samples for training, 2,000 samples for validation and 2,000 samples for testing. Figure 6 illustrates the frequency of occurrence in the training sets of CelebaAMask-HD and LaPa. Label distribution in LaPa is more uniform compared to CelebAMask-HQ. Notably, the frequency of samples for eye glasses, hats, and necklaces is lower than samples for other common categories, posing a greater challenge for the face parsing task.

Figure 6 Label occurrence frequency of CelebAMask-HD and LaPa train sets (see online version for colours)



Implementation details

Our method, built on PyTorch Vision 1.12.1 with Ubuntu 20.04, is trained from the ground up on four 2080Ti GPU cards, with no pre-training involved. The input images are sized at 448 × 448 for both the training and testing phases. Our model is exclusively trained on the train set, with evaluations and tests carried out on the validation and test sets. For data augmentation during training, we employ random rotation within (–20°, 20°), random shearing, random contrast adjustment, and random colour jittering. The mini-batch size is set at 8 per GPU card, totalling 32, and we enhance this further to a total batch size of 160 through gradient accumulation. The model is trained for 200 epochs, with the learning rate initially set at 5e-4 and gradually reduced to 1e-5. For transformer structure, we set patch size and window size for window attention to 4 and 7. Embedding dimensions in four hierarchical stages are set to 96, 192, 384 and 768. Feature map sizes in four hierarchical stages are 56, 28, 14 and 7. And the heads number of four hierarchical stages are set to 1, 3, 6, 12.

Evaluation metrics

In this paper, to maintain consistency in comparisons with prior research, we calculate the mean F1 score for both the CelebAMask-HD and LaPa datasets, considering all facial components and excluding the background.

$$precision = \frac{TP}{TP + FP} \tag{10}$$

$$recall = \frac{TP}{TP + FN} \tag{11}$$

$$F1 = \frac{2 * precision * recall}{precision + recall} \tag{12}$$

Table 1 Input size, parameters and MACs for several models calculated by using pytorch-opcounter

Setting	*Input size*	*Param*	*MACs*
Ours	448 × 448	29.15 M	22.05 G
DML-CSR (Zheng et al., 2022)	473 × 473	67.94 M	252.58 G
AGRNet (Te et al., 2021)	473 × 473	66.42 M	205.14 G
EAGR (Te et al., 2020)	473 × 473	66.61 M	236.41 G
EHANet (Luo et al., 2020)	256 × 256	43.90 M	11.50 G

4.1 Comparison with advanced model

In this paper, we compared proposed method with existing SOTA methods (Zhao et al., 2017; Liu et al., 2020; Te et al., 2020; Zheng et al., 2022; Lee et al., 2020; Wei et al., 2019; Luo et al., 2020; Te et al., 2021; Lin et al., 2021) on CelebAMask-HQ and LaPa datasets. Tab.3 and Tab.2 present a statistical comparison, highlighting the notable advancements achieved by our methods. The FP-transformer attains mean F1 scores of 87.1% on CelebAMask-HQ and 92.6% on LaPa, respectively. From Tables 2 and 3, it can be observed that the FP-transformer consistently outperforms other methods in most categories. Additionally, since the Lapa dataset has more balanced label distributions and fewer categories, models trained on the Lapa dataset achieve higher F1 scores for categories such as nose, mouth, and glasses. In contrast, models trained on the CelebAMask-HQ dataset exhibit lower F1 scores for these categories. Without using data augmentation and class weighting strategies, categories with fewer samples, such as glasses, necklaces, and earrings, tend to have relatively lower F1 scores. However, our model still demonstrates better performance in these categories compared to previous algorithms. Additionally, Tab.1 compares the input size, parameters, and MACs counts of various methods using pytorch-opcounter. This comparison reveals that the FP-transformer requires fewer resources and lowers computational complexity than other state-of-the-art methods. Figure 7 displays the visual outcomes of our methods compared to DML-CSR by Zheng et al. (2022), which previously showed the best performance on LaPa and CelebAMask-HQ. In Figure 1 our algorithm demonstrates superior performance on samples where other algorithms fail, showcasing its ability to segment faces more accurately. The first row presents the original RGB images of faces. The second row depicts the ground truth. In the third row, we see the unsuccessful results from Zhang et al.'s method in scenarios with complex backgrounds, low lighting, and facial obstructions. The fourth row illustrates the outcomes of our method, showcasing improved performance on these challenging samples. Our LS attention ensures the extraction of relationships for long-term features while preserving local features, enhancing the robustness of our method in challenging situations. It is evident that our method demonstrates greater robustness in complex scenarios involving multiple people, varied lighting, angles, and facial expressions. Notably, in images with sophisticated backgrounds or facial obstructions, our method appears more adept at differentiating facial features from the surrounding elements.

Table 2 Comparison in LaPa with SOTA methods in mean F1

Method	*Sk*	*Ha*	*LE*	*RE*	*UL*	*Mo*	*LL*	*No*	*LB*	*RB*	*Mean F1*
Zhao et al. (2017)	93.5	94.1	86.3	86	83.6	86.9	84.7	94.8	86.8	86.9	88.4
Liu et al. (2020)	97.2	96.3	88.1	88	84.4	87.6	85.7	95.5	87.7	87.6	89.8
Te et al. (2020)	97.3	96.2	89.5	90	88.1	90	89	97.1	86.5	87	91.1
Zheng et al. (2022)	97.6	96.4	*91.8*	*91.5*	88	90.5	89.9	*97.3*	90.4	90.4	92.4
Te et al. (2021)	97.7	96.5	91.6	91.1	88.5	90.7	90.1	97.3	89.9	90	92.3
Lin et al. (2021)	97.8	96.5	91.5	90.9	*88.7*	90.5	*90.5*	96.9	90.1	89.1	92.5
Our method	*97.9*	*96.8*	91.1	90.9	88.0	*91.0*	90.0	97.2	*91.1*	*90.9*	*92.6*

Notes: SK denotes skin. Ha denotes hair. LE and RE mean left and right eye. LB and RB mean left and right brow. Mo denotes mouth. UL and LL mean upper and lower lip. No denotes nose.

Table 3 Comparison in CelebAMask-HQ with SOTA methods in mean F1

Method	*Skin Mou*	*Nose UL*	*Gl LL*	*LE Hair*	*RE Hat*	*LB ER*	*RB NL*	*LR Nk*	*RR Cl*	*Mean F1*
Zhao et al. (2017)	94.8	90.3	75.8	79.9	80.1	77.3	78.0	75.6	73.1	76.2
	89.8	87.1	88.8	90.4	58.2	65.7	19.4	82.7	64.2	
Lee et al. (2020)	95.5	85.6	92.9	84.3	85.2	81.4	81.2	84.9	83.1	80.3
	63.4	88.9	90.1	86.6	91.3	63.2	26.1	92.8	68.3	
Wei et al. (2019)	96.4	91.9	89.5	87.1	85.0	80.8	82.5	84.1	83.3	82.1
	90.6	87.9	91.0	91.1	83.9	65.4	17.8	88.1	80.6	
Luo et al. (2020)	96.0	93.7	90.6	86.2	86.5	83.2	83.1	86.5	84.1	84.0
	93.8	88.6	90.3	93.9	85.9	67.8	30.1	88.8	83.5	
Te et al. (2020)	96.2	94.0	92.3	88.6	88.7	85.7	85.2	88.0	85.7	85.1
	95.0	88.9	91.2	94.9	87.6	68.3	27.6	89.4	85.3	
Te et al. (2021)	96.5	93.9	91.8	88.7	89.1	85.5	85.6	88.1	88.7	85.5
	92.0	89.1	91.1	95.2	87.2	69.6	32.8	89.9	84.9	
Zheng, et al. (2022)	95.7	93.9	*92.6*	89.4	89.6	85.5	85.7	88.3	88.2	86.1
	91.8	87.4	9.01	94.5	*88.5*	*71.4*	40.6	89.6	85.7	
Our method	*96.7*	*94.3*	92.5	*90.4*	*90.4*	*86.7*	*86.7*	*89.1*	*88.8*	*87.1*
	92.9	*89.9*	*91.5*	*95.6*	88.1	69.5	*46.5*	*91.3*	*86.8*	

Notes: GL denotes glasses. LE and RE mean left and right eye. LB and RB mean left and right brow. LR and RR denote left and right ear. Mou denotes mouth. UL and LL mean upper and lower lip. ER means earring. NL mean necklace. NK means neck. CL means cloth.

Figure 7 Visual comparison between ours method and existing SOTA method (see online version for colours)

4.2 Ablation study

Analysis of improvement. To demonstrate the impact of various components, we sequentially integrate our modules into the baseline model, evolving from a pure swin transformer to our proposed model. We begin with a pure swin transformer, comprising 1 Swin blocks at each stage with LayerNorm, as the baseline. This is followed by the incremental addition of convolution operations (ceasing the use of shifted windows), feature fusion, BLN, and our long-short attention (LS attention) block. As shown in Table 4, the baseline model utilises the Swin block without the shifted window mechanism. The performance improves progressively with the addition of convolutional operations, feature fusion, BLN, and LS Attention. Specifically, the full model, which incorporates all these components, achieves the highest results on the CelebAMask-HQ dataset, with a mean F1 score of 87.09 (an improvement of +2.86) and a mIoU of 79.40 (an improvement of +2.95) compared to the baseline. These results demonstrate that the integration of these features effectively enhances the model's performance in facial parsing tasks.

Table 4 Ablation study

Baseline	*Conv operation*	*Feature fusion*	*BLN*	*LS attention*	*CelebAMask-HQ*	
					Mean F1	*mIoU*
✓					84.23	76.39
✓			✓		84.43(+0.20)	76.95(+0.48)
✓	✓				84.67(+0.43)	77.21(+0.72)
✓	✓	✓			85.86(+1.63)	78.25(+1.76)
✓	✓	✓	✓		86.22(+1.99)	78.67(+2.18)
✓	✓	✓	✓	✓	*87.09(+2.86)*	*79.40 (+2.95)*

To further explore the different between LS attention and traditional attention block, we draw the GradCam (Selvaraju et al., 2017) visual explanations for the traditional attention block, LS attention block, self-attention branch and CNN branch in LS attention block. The brighter a pixel is, the greater its relevance to a specific category. Figure 8 presents a

series of Grad-CAM visualisations that illustrate the pixel-level contribution towards identifying the nose region in facial recognition tasks using various attention mechanisms. Specifically, Figure 8 displays the heatmap generated by a traditional attention mechanism, revealing where the model focuses when determining the nose area. The CNN branch within the LS attention block targets local features, primarily highlighting the edges of various facial parts, thereby aiding the LS attention block in more accurately predicting nose components. Meanwhile, the self-attention branch in LS attention has a broader focus, encompassing several larger areas. This feature enables the LS attention block to better extract relationships between different facial parts. Following the integration of information from both branches, LS attention achieves more robust and precise detection of the nose. This not only enhances the accuracy of its positional information but also improves the delineation of the nose's shape and edge.

Figure 8 Grad-Cam images comparison between traditional attention and LS attention (see online version for colours)

The traditional column represents the Grad-CAM visualisations derived from the traditional self-attention block, while the LS Attention column corresponds to the outputs of the Long-Short Range Attention Block. The attention branch and CNN branch columns display the Grad-CAM visualisations of the respective branches. It can be observed that the attention branch focuses more on global information, whereas the CNN branch captures facial edge details. These complementary features enable the LS Attention Block to produce outputs with sharper and more distinct partition boundaries.

5 Conclusions

This paper presents the FP-transformer, a comprehensive end-to-end face parsing model based on transformer. Extensive experiments conducted on CelebAMask-HQ and LaPa demonstrate the effectiveness and precision of our proposed method. The results consistently indicate that the FP-transformer substantially enhances face parsing performance, largely attributable to our proposed modules. Particularly, the Long-short

attention block introduced in this model effectively augments the local feature extraction capabilities of conventional attention blocks, enabling transformer-based models to more accurately predict the edges of different facial parts.

Acknowledgements

This work was supported by Science and Technology Department of Sichuan Province (No.2022YFO0056).

References

Ba, J.L., Kiros, J.R. and Hinton, G.E. (2016) 'Layer normalization', *arXiv Preprint*, arXiv: 1607.06450.

Chen, C., Fan, Q. and Panda, R. (2021) 'CrossViT: cross-attention multi-scale vision transformer for image classification', *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp.347–356.

Cheng, B., Misra, I., Schwing, A.G., Kirillov, A. and Girdhar, R. (2022) 'Masked-attention mask transformer for universal image segmentation', in *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pp.1290–1299.

Chu, X., Tian, Z., Zhang, B., Wang, X. and Shen, C. (2021) 'Conditional positional encodings for vision transformers', *International Conference on Learning Representations*.

Dai, X., Chen, Y., Yang, J., Zhang, P., Yuan, L. and Zhang, L. (2021a) 'Dynamic detr: end-to-end object detection with dynamic attention', in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp.2988–2997.

Dai, Y., Gieseke, F., Oehmcke, S., Wu, Y. and Barnard, K. (2021b) 'Attentional feature fusion', in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp.3560–3569.

Dai, Z., Cai, B., Lin, Y. and Chen, J. (2021c) 'Up-detr: unsupervised pre-training for object detection with transformers', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.1601–1610.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J. and Houlsby, N. (2020) 'An image is worth 16x16 words: transformers for image recognition at scale', *arXiv Preprint*, arXiv: abs/2010.11929.

Ge, Y., Liu, H., Du, J., Li, Z. and Wei, Y. (2023) 'Masked face recognition with convolutional visual self-attention network', *Neurocomputing*, Vol. 518, pp.496–506.

Guo, T., Kim, Y., Zhang, H., Qian, D., Yoo, B., Xu, J., Zou, D., Han, J. J. and Choi, C. (2018) 'Residual encoder decoder network and adaptive prior for face parsing', *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32, No. 1, https://doi.org/10.1609/aaai.v32i1.12268.

He, K., Zhang, X., Ren, S. and Sun, J. (2015) 'Deep residual learning for image recognition', *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.770–778.

Jackson, A.S., Valstar, M. and Tzimiropoulos, G. (2016) 'A CNN cascade for landmark guided semantic part segmentation', in *Computer Vision-ECCV 2016 Workshops*, Amsterdam, the Netherlands, 8–10 15–16 October, proceedings, Part III, Springer, Vol. 14, pp.143–155.

Jayaraman, U., Gupta, P., Gupta, S., Arora, G. and Tiwari, K. (2020) 'Recent development in face recognition', *Neurocomputing*, Vol. 408, pp.231–245.

Ji, Z. and Zhong, X. (2024) 'Bidirectional attention network for real-time segmentation of forest fires based on UAV images', *Int. J. Inf. Commun. Technol.*, Vol. 25, No. 6, pp.38–51.

Lee, C. H., Liu, Z., Wu, L. and Luo, P. (2020) 'Maskgan: towards diverse and interactive facial image manipulation', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.5549–5558.

Lin, F., Yuan, J., Wu, S., Wang, F. and Wang, Z. (2023) 'UniNeXt: exploring a unified architecture for vision recognition', *arXiv Preprint*, arXiv: 2304.13700.

Lin, J., Yang, H., Chen, D., Zeng, M., Wen, F. and Yuan, L. (2019) 'Face parsing with RoI tanh-warping', in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.5647–5656.

Lin, T.Y., Goyal, P., Girshick, R., He, K. and Dollár, P. (2016) 'Focal loss for dense object detection', in Ba, J.L., Kiros, J.R. and Hinton, G.E.: *Layer Normalization*, *arXiv Preprint*, arXiv: 1607.06450.

Lin, Y., Shen, J., Wang, Y. and Pantic, M. (2021) 'Roi tanh-polar transformer network for face parsing in the wild', *Image and Vision Computing*, Vol. 112, p.104190.

Liu, S., Yang, J., Huang, C. and Yang, M.H. (2015) 'Multi-objective convolutional learning for face labeling', in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.3451–3459.

Liu, Y., Shi, H., Shen, H., Si, Y., Wang, X. and Mei, T. (2020) 'A new dataset and boundary-attention semantic segmentation for face parsing', *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34, No. 7, pp.11637–11644.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S. and Guo, B. (2021) 'Swin transformer: hierarchical vision transformer using shifted windows', *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp.9992–10002.

Liu, S., Shi, J., Liang, J. and Yang, M.H. (2017) 'Face parsing via recurrent propagation', *arXiv Preprint*, arXiv: 1708.01936.

Luo, L., Xue, D. and Feng, X. (2020) 'Ehanet: an effective hierarchical aggregation network for face parsing', *Applied Sciences*, Vol. 10, No. 9, p.3135.

Luo, P., Wang, X. and Tang, X. (2012) 'Hierarchical face parsing via deep learning', in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp.2480–2487.

Mallick, S., Paul, J. and Sil, J. (2023) 'Response fusion attention u-ConvNext for accurate segmentation of optic disc and optic cup', *Neurocomputing*, Vol. 559, p.126798.

Minaee, S., Boykov, Y., Porikli, F.M., Plaza, A.J., Kehtarnavaz, N. and Terzopoulos, D. (2020) 'Image segmentation using deep learning: a survey', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 44, No. 7, pp.3523–3542.

Ronneberger, O., Fischer, P. and Brox, T. (2015) 'U-net: convolutional networks for biomedical image segmentation', in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015: 18th International Conference*, Munich, Germany, 5–9 October, Springer, Proceedings, Part III, Vol. 18, pp.234–241.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. and Batra, D. (2017) 'Grad-cam: visual explanations from deep networks via gradient-based localization', in *Proceedings of the IEEE International Conference on Computer Vision*, pp.618–626.

Siddique, N., Paheding, S., Elkin, C.P. and Devabhaktuni, V. (2021) 'U-net and its variants for medical image segmentation: a review of theory and applications', *IEEE Access*, Vol. 9, pp.82031–82057.

Song, Y., Tang, H., Meng, F., Wang, C., Wu, M., Shu, Z. and Tong, G. (2022) 'A transformer-based low-resolution face recognition method via on-and-offline knowledge distillation', *Neurocomputing*, Vol. 509, pp.193–205.

Te, G., Hu, W., Liu, Y., Shi, H. and Mei, T. (2021) 'Agrnet: adaptive graph representation learning and reasoning for face parsing', *IEEE Transactions on Image Processing*, Vol. 30, pp.8236–8250.

Te, G., Liu, Y., Hu, W., Shi, H. and Mei, T. (2020) 'Edge-aware graph representation learning and reasoning for face parsing', *arXiv Preprint*, arXiv: abs/2007.11240.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A. and Jégou, H. (2021) 'Training data-efficient image transformers & distillation through attention', in *International Conference on Machine Learning*, PMLR, pp.10347–10357.

Umirzakova, S. and Whangbo, T.K. (2022) 'Detailed feature extraction network-based fine-grained face segmentation', *Knowledge-Based Systems*, Vol. 250, p.109036.

Wang, W., Xie, E., Li, X., Fan, D., Song, K., Liang, D., Lu, T., Luo, P. and Shao, L. (2021) 'Pyramid vision transformer: a versatile backbone for dense prediction without convolutions', *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp.548–558.

Wei, Z., Liu, S., Sun, Y. and Ling, H. (2019) 'Accurate facial image parsing at real-time speed', *IEEE Transactions on Image Processing*, Vol. 28 No. 9, pp.4659–4670.

Woo, S., Debnath, S., Hu, R., Chen, X., Liu, Z., Kweon, I. and Xie, S. (2023) 'ConvNeXt V2: co-designing and scaling ConvNets with masked autoencoders', *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.16133–16142.

Wood, E., Baltrušaitis, T., Hewitt, C., Dziadzio, S., Cashman, T.J. and Shotton, J. (2021) 'Fake it till you make it: Face analysis in the wild using synthetic data alone', in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp.3681–3691.

Wu, H., Xiao, B., Codella, N.C., Liu, M., Dai, X., Yuan, L. and Zhang, L. (2021) 'CvT: introducing convolutions to vision transformers', *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp.22–31.

Wu, Y. (2024) 'Attention is all you need for boosting graph convolutional neural network', *arXiv Preprint*, arXiv: abs/2403.15419.

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M. and Luo, P. (2021) 'SegFormer: simple and efficient design for semantic segmentation with transformers', *Advances in Neural Information Processing Systems*, Vol. 34, pp.12077–12090.

Yu, Y., Wang, C., Fu, Q., Kou, R., Huang, F., Yang, B., Yang, T. and Gao, M. (2023) 'Techniques and challenges of image segmentation: a review', *Electronics*, Vol. 12, No. 5, p.1199.

Yuan, L., Chen, Y., Wang, T., Yu, W., Shi, Y., Tay, F.E., Feng, J. and Yan, S. (2021) 'Tokens-to-token ViT: training vision transformers from scratch on ImageNet', *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp.538–547.

Zhang, S., Tong, H., Xu, J. and Maciejewski, R. (2019) 'Graph convolutional networks: a comprehensive review', *Computational Social Networks*, Vol. 6, No. 1, pp.1–23.

Zhang, W., Tan, Q., Li, P., Zhang, Q. and Wang, R. (2023) 'Cross-modal transformer with language query for referring image segmentation', *Neurocomputing*, Vol. 536, pp.191–205.

Zhao, H., Shi, J., Qi, X., Wang, X. and Jia, J. (2017) 'Pyramid scene parsing network', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.2881–2890.

Zhao, Y., Zhao, C. and Song, X. (2023) 'Visual segmentation of the diagnosis image of pulmonary nodules with vascular adhesion based on convolution neural network', *Int. J. Inf. Commun. Technol.*, Vol. 22, No. 2, pp.147–161.

Zheng, Q., Deng, J., Zhu, Z., Li, Y. and Zafeiriou, S. (2022) 'Decoupled multi-task learning with cyclical self-regulation for face parsing', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.4156–4165.

Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P.H. and Zhang, L. (2020) 'Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers', *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.6877–6886.

Zheng, Y., Yang, H., Zhang, T., Bao, J., Chen, D., Huang, Y., Yuan, L., Chen, D., Zeng, M. and Wen, F. (2021) 'General facial representation learning in a visual-linguistic manner', *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.18676–18688.

Zhou, L., Liu, Z. and He, X. (2017) 'Face parsing via a fully-convolutional continuous CRF neural network', *arXiv Preprint*, arXiv: 1708.03736.

Zhou, Y., Hu, X. and Zhang, B. (2015) 'Interlinked convolutional neural networks for face parsing', in *Advances in Neural Networks – ISNN 2015: 12th international symposium on neural networks*, ISNN 2015, Jeju, South Korea, 15–18 October, Springer, Vol. 12, pp.222–231.

Zhu, X., Su, W., Lu, L., Li, B., Wang, X. and Dai, J. (2020) 'Deformable detr: deformable transformers for end-to-end object detection', *arXiv Preprint*, arXiv: 2010.04159.