ABANet: Attention Boundary-Aware Network for image segmentation

Sadjad Rezvani[a], Mansoor Fateh[a,*], Hossein Khosravi[b]

[a]*Faculty of Computer Engineering, Shahrood University of Technology, Daneshgah Blvd., Shahrood, iran.*
[b]*Faculty of Electrical Engineering, Shahrood University of Technology, Daneshgah Blvd., Shahrood, iran.*

Abstract

Deep learning techniques have attained substantial progress in various face-related tasks, such as face recognition, face inpainting, and facial expression recognition. To prevent infection or the spread of the virus, wearing of masks in public places has been mandated following the COVID-19 epidemic, which has led to face occlusion and posed significant challenges for face recognition systems. Most prominent masked face recognition solutions rely on mask segmentation tasks. Therefore, segmentation can be used to mitigate the negative impacts of wearing a mask and improve recognition accuracy. Mask region segmentation suffers from two main problems: there is no standard type of masks that people wear, they come in different colors and designs, and there is no publicly available masked face dataset with appropriate ground truth for the mask region. In order to address these issues, we propose an encoder-decoder framework that utilizes a boundary-aware attention network combined with a new hybrid loss to provide a map, patch, and pixel-level supervision. We also introduce a dataset called MFSD, with 11601 images and 12758 masked faces for masked face segmentation. Furthermore, we compare the performance of different cutting-edge deep learning semantic segmentation models on the presented dataset. Experimental results on the MSFD dataset reveal that the suggested approach outperforms state-of-the-art, algorithms with 97.623% accuracy, 93.814% IoU, and 96.817% F1-score rate. Our dataset of masked faces with mask region labels and source code will be available online.

Keywords: deep learning, semantic segmentation, Attention Gate (AG), Masked Face Segmentation Dataset

1. Introduction

The development of deep convolutional neural networks (CNNs) has attracted considerable interest in face recognition in recent years [1]. Face recognition systems have been widely used in a variety of applications, including visual surveillance [2], automated border control [3], education systems [4] and healthcare [5]. Face recognition technology needs to be more efficient while confronting obstacles such as varying illumination [6], low resolution [7], different pose [8], expression change [9] and occlusion [10, 11, 12].

Occlusion in face recognition refers to the partial or complete obstruction of facial features by objects, accessories, or other elements. The obstructed areas may contain crucial information for accurate identification. Therefore, occlusion is a significant challenge for facial recognition algorithms. These obstructions can include accessories like sunglasses, hats, or face masks. Covering or hiding parts of the face reduces the algorithm's ability to extract meaningful facial feature, which leads to a decrease in recognition accuracy.

In the context of the COVID-19 pandemic, the widespread use of face masks has become a common form of occlusion. Besides, masks allow fraudsters and thieves to steal and commit crimes without identification. Face masks cover a substantial portion of the lower face, including the mouth and nose. These are crucial regions for facial recognition algorithms as they often rely on features like the nose shape, mouth structure, and the distance between these features for accurate identification. In fact, Due to the occlusion of human faces by masks, face recognition systems are faced with a serious challenge since approximately half of the biometric information is lost [13]. Therefore, it is particularly essential to address the challenges posed by mask-wearing to improve facial recognition performance.

In general, the issue of masked face recognition systems mainly consists of four fundamental steps. (1) masked face detection, (2) Pre-processing of the masked face image, (3) robust feature extraction, and (4) classification. Masked face recognition considers a heavy occlusion problem due to the mask covering almost half of the face. More specifically, approximately half of the critical facial semantics are lost, suggesting one's identity is complicated. As a result, pre-processing methods play an essential role in extracting robust features. Two approaches for pre-processing have been proposed [14]: representation and reconstruction.

In the representation approach, the mask is first seg-

*Corresponding author.
Email addresses: sadjadrezvani@shahroodut.ac.ir (Sadjad Rezvani), Mansoor_fateh@shahroodut.ac.ir (Mansoor Fateh), hosseinkhosravi@shahroodut.ac.ir (Hossein Khosravi)

mented and entirely excluded from the feature extraction. This method recognizes the face based on the unmasked facial region [14]. In contrast, reconstruction techniques tackle the occlusion problem in image space by restoring facial parts hidden behind the mask to resemble the original face image. This recovery process is contingent upon having a binary segmented region of mask, as emphasized in previous works [15, 16, 17]. Notably, reconstruction approaches have demonstrated enhanced recognition performance, particularly for faces with substantial occlusions like masks [18].

So The most prominent masked face recognition solutions rely on mask segmentation tasks. As a result, segmentation can be used to reduce the negative effects of wearing a mask and boost recognition accuracy. [19].

Mask segmentation is challenging for the following reasons: (1) There is no standard type of mask that people wear; it comes in different colors and designs, and (2) there is no masked dataset with appropriate ground truth for the mask regions to be used in the training phase. To make a step forward in masked face segmentation and address the above challenges, we collect masked face segmentation datasets. Also, a novel convolutional network framework is proposed, which is made up of encoders and decoders that can effectively segment the salient object regions. Specifically, we adopted the attention gate (AG) mechanism [20] to enhance the network's capacity for learning. An attention gate module was implemented in the skip connection part to identify salient feature regions further.

Our three main contributions are as follows:

(1) Design and development of an ABANet for highly accurate image segmentation.

(2) Preparing a dataset for masked face segmentation, denoted as MFSD, to improve masked face-related task performance. The dataset consists of 12758 Internet images, in which 11601 masked human faces are manually segmented.

(3) A comprehensive review of the proposed model was conducted, along with a comparison with 10 other segmentation networks

The structure of this paper is as follows: Section 2 provides a review of recent related research undertaken by various researchers; Section 3 presents the proposed dataset; Section 4 explains the model architecture; and Section 5 presents an experiment performed on the MFSD dataset to demonstrate the effectiveness of the proposed approach; Section 6 concludes the paper and states the future work.

2. Related work

A new dataset and a novel mask segmentation model are the major contributions of this work. Therefore, we briefly review the related works in three aspects. In the beginning, we review some mask region segmentation-related task approaches. Next, we present attention mechanism methods. Finally, we review the mask face datasets.

2.1. Mask Region Segmentation

Masked face recognition (MFR) is a major task, particularly during the global outbreak of COVID-19. One intuitive approach to recognizing faces under a mask can be a segmentation-based strategy that first detects the occluded region part and uses only the non-occluded part or intends to recover an occlusion-free face from the occluded mask face.

[19], Used a modified version of Unet to process masked face images, and then a binary segmented mask region is used to inpaint fine facial detail while maintaining the global coherence of the face. IAMGAN was proposed by [21] as a solution to the issue of insufficient data as well as an improvement to the discriminative capacity of MFR models. Training of a segmentation network generates the mask region. [22], proposed a new 3D reconstruction-based method to remove masks from face images. The model uses residual blocks to segment masks. In [23], segmentation of mask areas from masked face images is achieved through the utilization of the grab-cut method [24]. These studies demonstrate good results in the synthetic masked face dataset. These models are unable to generate an appropriate segmentation map of the mask object when applied to real mask pictures that have a variety of forms and structures.

2.2. Attention-based Methods

The human perception process inspires the attention mechanism to allow the system to ignore irrelevant information and focus on the most important local features. Attention mechanisms have succeeded in many visual tasks, including semantic segmentation [25, 26] , object detection [27], image classification [28], image generation, and self-supervised learning.

FocusNet [29] is a fully convolutional network that incorporates attention into segmenting medical images based on feature maps generated by a convolutional autoencoder. [30] Developed a dual attention network that combines position attention with channel attention to segment scenes based on interdependent channel maps. Zhang *et al.* [23] presented a masked face recognition system called AMaskNet, which includes a feature extractor and a contribution estimator module that utilizes attention-awareness. The contribution estimator includes attention mechanisms for both spatial and channel dimensions. Li *et al.* [31] developed a method for mask face recognition that cropped the input image to focus on the region around the eyes. As part of the attention-based component, they used a convolutional block attention module (CBAM) [32].

2.3. Masked Face Datasets

Inspired by the COVID-19 pandemic, people wear face masks. In such a scenario, a large dataset of masked faces is essential for deep-learning models to detect people wearing masks. Although various face databases have become publicly available in recent years, new large-scale datasets

are required to identify masked faces. We first break down masked face datasets into real or simulated.

2.3.1. Real Datasets

MAFA dataset [33] contains around 35806 images of masked faces with a diverse orientation of faces and a degree of occlusion. The dataset was generated by collecting images from various sources on the internet. MFDD(Masked Face Detection Dataset) [34] is an insightful dataset consisting of around 24500 masked face images. It will be highly beneficial to train models for masked face detection. Typically, gallery images are biased toward the Chinese face.

RMFRD(Real-World Masked Face Recognition Dataset) [34] includes 5,000 photos of 525 individuals wearing masks and 90,000 photos of the same 525 individuals without masks. With the limited number of masked faces, this dataset is not appropriate for detection tasks because it does not provide the coordinates of a rectangle outlining the mask area, but it is helpful for face recognition.

2.3.2. Simulated Datasets

Wang et al. [34] created a simulated masked face dataset that included 500,000 face photos from 10,000 different people. The photos included in the collection were from two different datasets: LFW [35] and Webface [36].

Based on the Flickr Faces HQ (FFHQ) [37] dataset, MaskedFace-Net [38] comprises 137,016 photos of correctly and badly worn masks. The collection includes both properly and incorrectly worn masked faces, as well as no masked faces.

An overview of the different datasets described in this section is illustrated in Table 1.

3. Proposed Dataset

During the covid-19 era wearing face masks posed new challenges to face-related tasks, including facial recognition, face inpainting, expression recognition, and object removal.

Mask region segmentation is a preliminary stage to tackle the occlusion issue corresponding to the face-related tasks [22]. Existing masked face datasets are not procedure binary segmentation maps because Segmenting mask regions manually is a time-consuming operation. As a result, existing unmasking methods [19, 22, 39, 40, 38, 18] synthesize training data by overlaying masks on existing face datasets. However, since these techniques rely on an artificially generated mask, their effects tend to seem unnatural. [41]. To address this issue, the masked face segmentation dataset(MFSD) provides the first public training dataset for the mask segmentation task. We present data collection process, mask annotation and Dataset Statistics:

3.1. Data collection process

A Python crawler script searches vast amounts of Internet data for front-face photographs of notable figures and related masked-face images. The data collection process involved the following steps:

- **Image Collection:** Over 25000 face images were collected using a python crawler tool from Internet resources like Google search engines and Instagram.
- **Automatic Filtering:** To ensure dataset quality, we employed the AIZoo face-mask detector [42] to automatically filter out irrelevant or non-masked images, resulting in a refined dataset.
- **Final Dataset:** After this filtering process, we retained a total of 11,601 images featuring 12,758 masked faces. Some illustrative examples of these masked faces are presented in Figure 1.

3.2. Mask annotation

We used the LabelMe toolbox [43] to manually segment the mask area of 2750 faces in our dataset. We train a model using the labeled samples of this dataset and predict the mask region of unlabeled samples. The model could not segment all the images correctly. As a result, in order to ensure the quality of the annotation, we used two experienced annotators to double-check the errors made by machines. The dataset creation process took approximately 6 months. Our datasets of masked faces with mask region labels will be available at https://github.com/sadjadrz/MFSD. Figure 2 shows some sample images and their corresponding labels from our proposed dataset.

3.3. Dataset statistics

In this section, we present comprehensive statistics about the MFSD dataset, including image sizes, face orientations, the distribution of masked faces per image, and mask types.

- **Image sizes:** After obtaining the images, those with a size lower than 100×100 were eliminated. The diversity in image sizes is visually represented in Figure. 3a, while the average size of the images in the dataset stands at 1024×1024 pixels.
- **Face orientations:** In Figure 3b, we present the data on face orientations. The majority of faces within the MFSD dataset are oriented directly forward, with only a limited number showing a left or right orientation. This diversity in orientations, including challenging cases like left-front and right-front faces, provides valuable opportunities to further evaluate their robustness.

Table 1: Major masked face datasets.

Dataset	Year	No. of Images	No. of masked faces	Real mask
MAFA [33]	2017	30811	35816	yes
MFDD [34]	2020	24771	-	yes
RMFRD [34]	2021	5000	5000	yes
SMFD [34]	2020	500000	500000	no
Masked Face-Net [38]	2021	137016	137016	no

Figure 1: Masked faces from our proposed dataset.

- **Number of masked faces per image:** The bar plot titled 'Number of masked faces per image' as depicted in Figure 3c, provides a visual representation of the distribution of the number of faces present within each image in the MFSD dataset. Each bar on the plot corresponds to a specific count of masked faces, and the height of each bar indicates the frequency of images with that particular count of faces.

  Upon examination of the plot, it becomes evident that the majority of images contain a single masked face, as indicated by the tallest bar. However, a non-negligible number of images feature multiple masked faces, represented by the bars to the right of the plot. This distribution sheds light on the prevalence of various face counts within the dataset and serves as valuable insight for understanding the dataset's composition.

- **Mask types:** The Mask Types pie chart, as presented in Figure 4, offers valuable insights into the distribution of different types of masks within the MFSD dataset. This chart illustrates the frequency with which various mask types are represented among the masked faces.

  As we can observe from the chart, different mask types are encountered in the dataset, each with its own prevalence. The chart segments are divided to represent specific mask categories, such as surgical masks, N95 masks, cloth masks, and others, including unique designer masks. The size of each segment corresponds to the proportion of images featuring that particular mask type.

  For instance, the dominant presence of cloth masks is clearly depicted by the largest segment in the chart. On the other hand, smaller segments represent less common mask types within the dataset.

  Understanding the distribution of mask types is essential in characterizing the dataset's composition and provides valuable context for researchers and practitioners working on tasks related to masked face recognition, segmentation, or analysis. This information aids in adapting algorithms and models to accommodate the variations in mask types, ensuring robustness and accuracy in real-world applications where different mask types may be encountered.

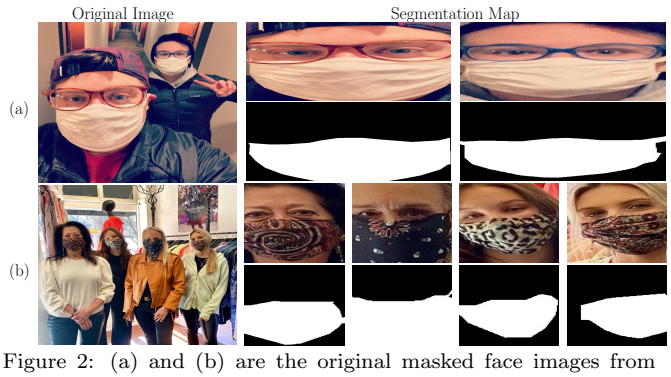
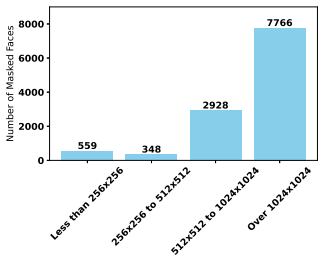
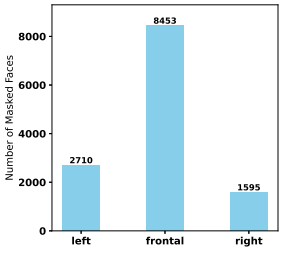


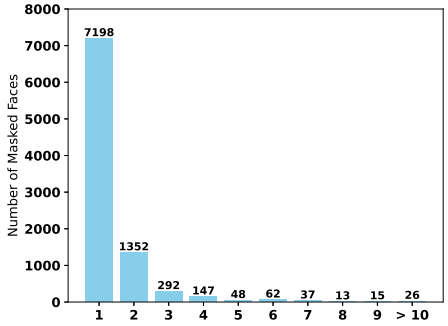
Figure 2: (a) and (b) are the original masked face images from MFSD; In the segmentation step, the top row is cropped faces, the second row is the face mask label of all faces in the original image.



(a) Image Sizes



(b) Face Orientation



(c) Number of masked faces per image

Figure 3: Dataset Statistics: (a) Image Sizes, (b) Face Orientation, (c) Occurrence Counts

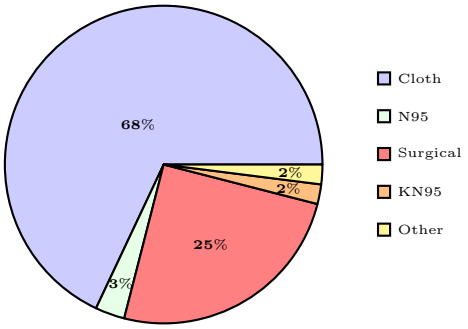


Figure 4: Distribution of Mask Types in the Dataset

4. Methodology

This section begins by providing an overview of the entire attention boundary-aware network (denoted as ABANet) and then describe the details of the network. After delving into the details of the network, we present an in-depth description of the attention mechanism, which plays a pivotal role in enhancing the network's performance. The network training loss is explained towards the end of this section.

4.1. Overview of the Proposed Network

The network architecture of our improved model is illustrated in Figure 5. Inspired by [44], our model comprises two stages: segmentation network and refinement network, which take a whole image as input and predict the saliency mask region in an end-to-end manner. Unlike the work [44], the segmentation subnet is an encoder-decoder network with an attention mechanism to identify the boundary between the mask and non-mask regions. The mask boundary was refined in the second stage using a fully convolutional U-Net like network.

4.2. Segmentation Network

The proposed segmentation network comprises three main parts: an encoder part, a decoder part, and an attention module. The encoded part was processed with six residual layers. At first, images are fed to 64 convolutional filters with a size of 3×3 and stride of 1. the first four stages are similar to ResNet-34 [45]. The fifth and sixth levels each include 512 filters and three fundamental res-blocks. In order to connect the encoder and the decoder, the bridge was constructed. It has three convolutional layers, each with 512 dilated 3×3 filters, for a total of 512 filters. Same as the encoder, the decoder has six stages. The first two stages consist of three convolution layers followed by a bilinear upsampling. The input of the first layer of the decoder comprises the concatenated feature maps from the corresponding stage in the encoder, acquired at the same spatial resolution as the current decoder stage during encoding. Additionally, it includes the feature maps from the last layer of the preceding decoder stage, embodying the high-level abstract features captured by the previous decoder stage. For other stages, the attention gate (AG) is added to our segmentation network to highlight salient features that are passed through the skip connections. By focusing on features closer to high-resolution feature maps, the AG optimizes the network's ability to capture intricate facial details and mask boundaries, all while maintaining computational efficiency. The Attention Gate can be summarized as follows figure 6. It can be observed that introducing an attention gate significantly enhances IoU. The output from the segmentation module's final stage is processed by a 3×3 convolution layer located at the decoder's conclusion before being forwarded to the refinement module.

4.3. Refinement Network

Similar to [44], the refinement network is designed using a residual encoder-decoder framework. Both the encoder and decoder are composed of four levels. Each level has a convolutional layer with 64 filters measuring 3×3, followed by normalization in batches and a ReLU activation function. Integrating high-level features with low-level features helps refine the object boundaries.

4.4. Attention Mechanism

The Attention Gate (AG) serves as a fundamental component within our segmentation network, significantly enhancing the precision of boundary-aware segmentation. As depicted in Figure 6, which illustrates the architecture of

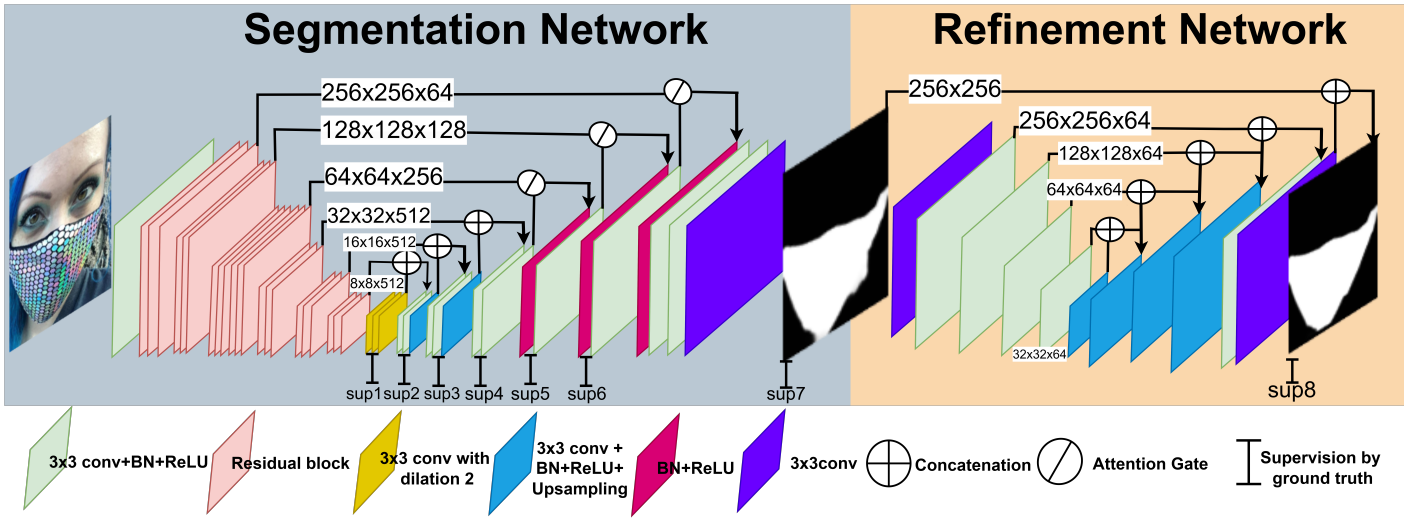


Figure 5: Architecture of our proposed ABANet network.

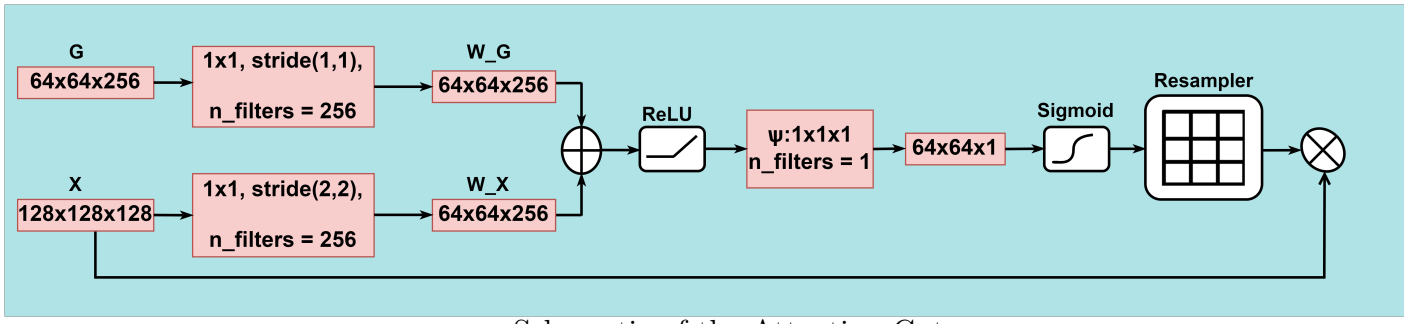


Figure 6: Schematic of the Attention Gate.

the Attention Gate, this mechanism plays a crucial role in focusing the model's attention on salient regions, ultimately leading to more accurate and detailed segmentation results.

4.4.1. Architecture of the Attention Gate

The Attention Gate's architecture, outlined in Figure 6, comprises distinct elements that collectively contribute to its functionality:

- **W_G and W_X:** As illustrated in Figure 5, the Attention Gate takes two sets of feature maps as input. These two feature maps are denoted as 'G' and 'X', as shown in Figure 6. The 'G' feature map originates from the decoder part of our network, while 'X' is obtained from the previous layer in the decoder. Both 'G' and 'X' are processed using convolutional layers, yielding 'W_G' and 'W_X' feature maps, respectively. Specifically, 'G' utilizes a convolutional layer with a kernel size of 1 and a stride of 1, while 'X' employs a convolutional layer with a kernel size of 1 and a stride of 2. These specific settings enable the layers to effectively extract essential information from both feature maps, which is integral to the attention mechanism's functionality.
- **Element-wise Sum and ReLU:** After processing, the results from 'W_G' and 'W_X' are element-wise summed and then passed through a Rectified Linear Unit (ReLU) activation function. This step facilitates the integration of information from both feature maps while introducing non-linearity into the attention mechanism.
- **Psi (ψ) Calculation:** Figure 6 also showcases how the combined result undergoes further processing through a convolutional layer. The application of a sigmoid activation function results in the generation of the attention map ψ, as illustrated in the figure. This attention map acts as a guiding mechanism for the model, directing its focus towards the most relevant spatial locations within 'X'.
- **Upsampling:** To ensure that the attention map aligns with the original input dimensions, Figure 6 indicates the application of an upsampling operation to ψ. This scaling operation restores the attention map to the same resolution as the input feature maps, aligning it with the spatial features of the original data.
- **Element-wise Multiplication with 'X':** After the

upsampling step, the output of the previous step is element-wise multiplied with This operation modulates the information in 'X' based on the attention map's guidance, resulting in a refined feature map that emphasizes salient spatial locations. This element-wise multiplication ensures that the final output retains the spatial structure and dimensions of the original 'X' feature map while emphasizing relevant information.

4.4.2. Integration into the Segmentation Network

Figure 6 provides an overview of the seamless integration of the Attention Gate into our segmentation network. As illustrated, multiple instances of the Attention Gate are strategically inserted at various points within the network architecture. This integration allows the model to adaptively attend to different scales and features, with a particular emphasis on salient regions and the intricate boundaries between mask and non-mask regions.

4.4.3. Choice of Feature-Wise Attention over Spatial Attention

In our ABANet design for mask face segmentation, we intentionally opted for feature-wise attention over spatial attention. This choice is underpinned by several key advantages:

- Enhanced Discriminative Power: Feature-wise attention allows the network to independently weigh the importance of each feature channel. This fine-grained control enables the network to emphasize specific discriminative features, such as facial edges and boundary information, essential for precise boundary-aware segmentation.

- Adaptability to Complex Structures: Mask face segmentation often involves intricate facial features and contours, which spatial attention mechanisms may struggle to capture adequately. Feature-wise attention excels in adaptability, dynamically emphasizing critical features, even when spatially distributed in a complex manner.

- Mitigation of Spatial Variability: Feature-wise attention is less sensitive to spatial variations in face orientation, scale, or position within the image. Operating on feature channels independently, it enhances the network's robustness across diverse input conditions frequently encountered in real-world scenarios.

In summary, Figure 5 provides a visual representation of the Attention Gate's architecture and its integration within our proposed segmentation network. This attention mechanism significantly enhances boundary-aware segmentation, leading to remarkable improvements in segmentation accuracy.

4.5. Hybrid Loss

We propose a hybrid loss function made of three components: focal loss, SSIM loss, and IoU loss, to provide superior regional segmentation and accurate boundaries. Our model has eight outputs, as shown in Figure 5. Thus, the whole segmentation loss can be described as:

$$Loss = \sum_{n=1}^{N} Loss^{(n)} \tag{1}$$

$$Loss^{(n)} = Loss_{focal}^{(n)} + Loss_{ssim}^{(n)} + Loss_{iou}^{(n)} \tag{2}$$

where, $Loss^{(n)}$ is the side output loss for the n-th output, and N is the overall number of outputs.

We leverage the focal loss as a pixel-level loss, to cope with low-confidence labels. It is defined as:

$$L_{binary-focal} = \begin{cases} -\alpha(1-y')^{\gamma}\log(y'), & y=1 \\ -(1-\alpha)(y')^{\gamma}\log(1-y'), & y=0 \end{cases} \tag{3}$$

where, y is the ground truth and y' is the prediction. To determine the value of α, we set up the ABANet when α={0.3, 0.5, 0.75 , 0.9} while $\gamma = 2$ and shows the quantitative comparison in Figure 7. In order to keep performance at an appropriate level, we set the $\alpha = 0.3$ on the basis of this comparison.

The SSIM loss is concentrated at the patch level, which can be utilized to capture the structural data to obtain a more detailed boundary prediction [46]. The SSIM loss definition is expressed in Eq. (4).

$$\text{SSIM} = 1 - \frac{(2\mu_x\mu_y + C_1) + (2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{4}$$

The covariance of x and y is represented by σ_{xy}. The mean and standard deviation of x and y are represented by μ_x, μ_y, σ_x, and σ_y, respectively. In order to prevent division by zero, we empirically fixed $\mathbf{C}_1 = 0.012$ and $\mathbf{C}_2 = 0.032$ in this paper [47].

The IoU loss is at map-level, thus focus on the foreground. IoU loss function is expressed by

$$\text{IoU} = 1 - \frac{\sum_{r=1}^{H}\sum_{c=1}^{W} S(r,c)G(r,c)}{\sum_{r=1}^{H}\sum_{c=1}^{W}[S(r,c) + G(r,c) - S(r,c)G(r,c)]} \tag{5}$$

where r and c represent the row and column indices, respectively. Specifically, r corresponds to the row index of a pixel in the image, and c corresponds to the column index of that pixel. G(r,c) represents the Ground Truth (GT) label of the pixel located at row r and column c, while S(r,c) represents the predicted probability of the segmented object at the same pixel location.

5. Experiments

In this section, we performed experiments to compare the performance of MFSD with 10 state-of-the-art deep learning models for segmentation, which were selected as baseline approaches for evaluation of our proposed network.

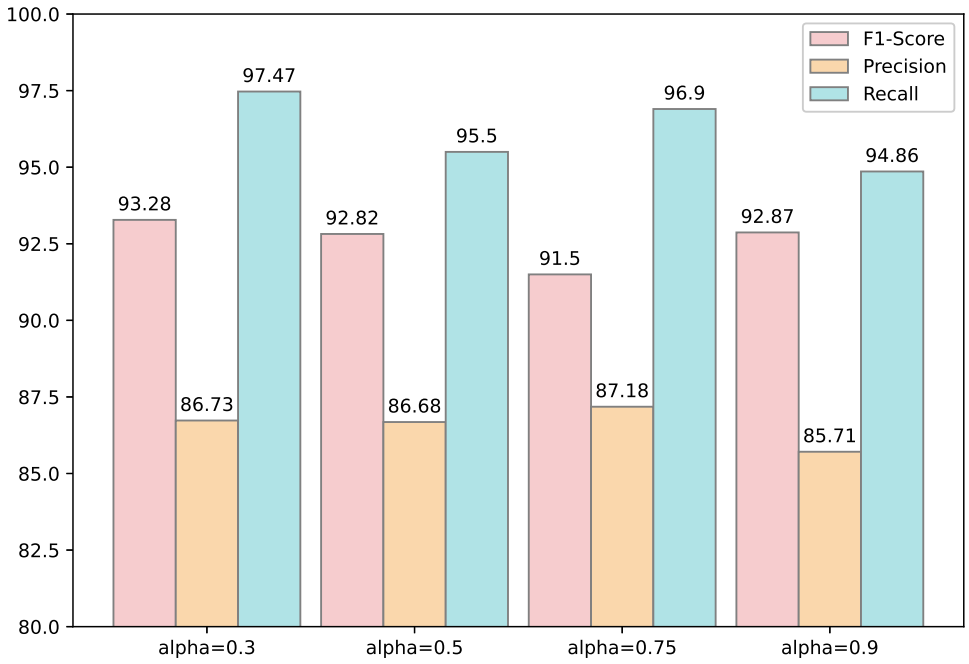


Figure 7: Quantitative comparison of α for focal loss.

5.1. Experimental Settings

In the training process, we split 12758 labeled images into three groups: the training set (8500 images), the validation set (2500 images), and the test set (1758 images). We train all models with ImageNet [48] pre-trained encoders. The calculation of the intersection-over-union (IoU) metric involves measuring the overlap between positive labels in a binary segmentation map. IoU is a frequently used standard statistic for assessing image segmentation techniques. All the networks are implemented in the segmentation models library based on Pytorch [49]. The training batch sizes were 8. We utilized the Adam optimizer [50] with a learning rate of 10^{-4} to optimize the parameters. As input, the model received images of 256 $\times$ 256 pixels.

We train models using multiple backbones, including vgg19, resnet50, and efficientnet.

5.2. Evaluation Metrics

We employed several metrics to comprehensively evaluate the performance of our network. These metrics include both traditional and specialized measures for semantic segmentation tasks. Here, we provide details on the five metrics used for evaluation:

5.2.1. Traditional Metrics

Intersection over Union (IoU) serves as the primary evaluation measure for semantic segmentation. Additionally, we calculate the following standard metrics [51]:

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

$$F1 - score = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \tag{8}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

$$IoU = \frac{TP}{TP + FP + FN} \tag{10}$$

5.2.2. Additional Metrics

In addition to the traditional metrics, we incorporated the following specialized metrics to provide a more comprehensive assessment:

$$MAE = \frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} |S(i,j) - G(i,j)| \tag{11}$$

where MAE represents the Mean Absolute Error, calculated as the average pixel-wise absolute difference between the predicted map S and the ground truth image G. W and H are the width and height of the saliency map and (i, j) denotes the pixel coordinates.

The S-measure, as described in the work by [52], quantifies the structural similarity between the predicted map

Table 2: A quantitative comparison of the results obtained on the MFSD test set using a variety of techniques for image segmentation.

Model	Backbone	IoU	F1-score	Precision	Recall	Acc
Unet	Efficient-b7	89.516	94.469	91.507	97.63	96.115
Unet	ResNet50	89.188	94.299	91.463	97.318	96.032
Unet	VGG19	89.208	94.302	91.28	97.533	96.012
Unet++	Efficient-b7	89.524	94.47	91.565	97.567	96.129
Unet++	ResNet50	88.526	93.928	90.952	97.106	95.678
Unet++	VGG19	89.242	94.329	91.602	97.225	95.998
MAnet	Efficient-b7	89.587	94.507	91.667	97.529	97.159
MAnet	ResNet50	89.61	94.21	91.218	97.405	95.934
MAnet	VGG19	89.242	94.339	91.45	97.418	95.999
Linknet	Efficient-b7	89.484	94.464	91.629	97.482	96.113
Linknet	ResNet50	89.011	94.202	91.427	97.152	95.904
Linknet	VGG19	89.064	94.239	91.652	96.978	95.905
DeepLabV3+	Efficient-b7	89.242	94.353	91.45	97.448	95.999
DeepLabV3+	ResNet50	89.463	94.46	91.46	97.665	96.083
PAN	ResNet50	89.312	92.967	91.522	97.435	96.033
PAN	VGG19	89.466	94.457	91.52	97.59	96.107
EINet	ResNet50	91.435	96.576	95.341	**97.845**	97.2
EU-Net	ResNet-34	88.834	94.11	91.09	97.352	95.875
DAD	ResNet50	93.24	96.598	95.57	97.649	97.39
BASNet		91.68	95.488	94.754	96.235	96.787
Ours		**93.814**	**96.817**	**97.164**	96.474	**97.623**

Table 3: Performance Comparison of the Top Five Models on additional metrics.

Model	MAE	F_β	S_m
EINet	0.026	0.955	0.957
EU-Net	0.035	0.915	0.872
DAD	0.027	0.957	0.96
BASNet	0.018	0.948	0.953
Ours	**0.014**	**0.971**	**0.968**

S and the ground truth image G by taking into account object-aware and region-aware aspects.

$$S\text{-measure} = \alpha \cdot S_o(S, G) + (1 - \alpha) \cdot S_r(S, G) \tag{12}$$

Where S_o represents the object-aware structural similarity, S_r represents the region-aware structural similarity, and α is a parameter. In accordance with prior work [52], we have chosen to set α to 0.5, which determines the balance between the importance of these structural similarities.

The F-measure represents a balanced combination of precision and recall, calculated based on the predicted maps and the ground truth images using the following formula:

$$F_\beta = \frac{(1 + \beta^2) \cdot \text{Precision} \cdot \text{Recall}}{\beta^2 \cdot \text{Precision} + \text{Recall}} \tag{13}$$

Here, we set the value of β^2 to 0.3, as recommended in [53], to emphasize the importance of precision.

Contrary to F_β and S_m, a lower MAE value signifies superior performance. We employed the implementations provided by [53] to calculate F_β, S_m, and MAE for our results.

5.3. Comparison with state-of-the-art models

On our masked face segmentation dataset, we conduct a comparison of our network with other state-of-the-art approaches using the same training configuration of 300 testing images. Segmentation models are Unet[54], Unet++[55], Linknet[56], MAnet[57], PAN[58], EINet[59], EU-Net[60], DAD[61], DeepLabV3+[62] and BASNet [44]. Also, we evaluate models on different backbones, such as ResNet-50, VGG19, and Efficient-b7.

Quantitative Evaluation: as shown in Table 2, the evaluation metrics were calculated and summarized For quantitative comparisons. The best scores are highlighted with bold. As can be observed from Table 2, our proposed

ABANet achieved 93.814, 96.817, 97.164, and 97.623 in terms of IoU, F1-score, precision, and accuracy, respectively, which were superior to those of other methods. The IoU and F1-Score of the proposed network were approximately 4% and 2%, respectively higher than those of the classical networks like Unet, Unet++, and Linknet, which implied that the proposed refinement network can help efficiently. Compared to newer methods like DAD, EU-Net, and EINet, the proposed network showed marginal improvements in the result, which indicates that hybrid loss is suitable for mask region segmentation. Compared with BASNet [44], a better comprehensive segmentation performance is achieved through our network, which is relatively 0.239% to 2.41% higher with different evaluation indicators. It is demonstrated that attention block capture more contextual information and is beneficial for segmentation tasks.

In order to gain a deeper understanding of the performance of the models evaluated in our study, we provide a qualitative comparison based on the performance metrics, specifically Mean Absolute Error (MAE), F_β score, and S-measure (S_m). The Table 3 displays the performance results of the top five models from Table 2. Notably, our model excels with the lowest MAE of 0.014, emphasizing superior pixel-wise performance. Furthermore, it achieves the highest F_β score of 0.971 and an impressive S-measure of 0.968, showcasing excellence in terms of boundary precision and structural similarity.

Qualitative Evaluation: Figure 8 provides qualitative comparisons between our ABANet and five other methods. As shown in Figure 8, the EU-Net, and U-Net++ methods are marginally less effective in segmenting mask regions, and the segmentation results appear noisy. As mentioned earlier, the state-of-the-art suffers from over-segmentation or under-segmentation when dealing with the mask covered by an object or hand. In contrast, our approach can precisely preserve the face mask's boundaries and structures. It can be observed from the red boxes that our network is more robust in describing boundaries and edges than BASNet.

To sum up, our ABANet has the ability to handle a wide variety of challenging cases, such as masks with different characteristics (e.g., color, size) and low resolution.

5.4. Comparison with Masked Face Segmentation Models

In this section, we provide a comprehensive comparison of our proposed ABANet with four existing masked face segmentation models. This comparison aims to evaluate ABANet performance in the context of masked face segmentation. While ABANet has demonstrated its superiority with overall image segmentation approaches, it is essential to assess its performance within the domain of masked face segmentation networks. This section focuses on the comparative analysis of out approach and other masked face segmentation models.

The training process aligns with the procedures detailed in the experimental settings section. To ensure the stability and consistency of our results, we performed a series of five trials for each experiment. For each trial, we randomly initialized the model weights and split the dataset into training, validation, and test sets. Subsequently, we computed the mean value for each metric across all trials and determined the corresponding standard deviation. This comprehensive approach enhances the robustness of our assessment of the model's performance, effectively accommodating variations that may arise due to differing random initializations or dataset splits.

DIN et. al [19] and Geng et. al [21] employ U-Net-based segmentation networks for mask region segmentation, while [63] utilizes the Mask-RCNN [64] model. In contrast, [23] employs the Grab-Cut[24] method for mask region segmentation in the context of masked face recognition.

As can be seen in Table 4, ABANet outperforms all the tested face mask region segmentation models in terms of the IoU, F1-score, F_β and MAE metrics. ABANet achieves IoU, F1-score, F_β and MAE scores of 93.4%, 96.5%, 96.8% and 1.4% which is 2%, 0.3%, 1.3% and 1.2% better than the scores of the second-best method (i.e., Sola and Gera [63]).

In addition to other performance metrics, we assess the model's inference speed, quantified in Frames Per Second (FPS). FPS is an indicator for evaluating the efficiency of image segmentation. A higher FPS value signifies a faster inference speed. When comparing our method to Zhang et al.'s [23] approach, although the FPS of this method is higher, our method exhibits a superior performance in the other four evaluation metrics, achieving an average improvement of 10 percent.

Figure 9 presents a selection of representative images obtained through the masked face segmentation methods described in Table 4, along with results from our proposed method. To conduct a detailed analysis, we deliberately chose challenging masks with unconventional shapes and designs. It can be seen from 9, ABANet can outperform other masked face segmentation network when encounter with different mask design and shape. Specifically, our network excels in the precise segmentation of entire mask regions while also providing clear and accurate boundary predictions

5.5. Ablation study

In this section, to further evaluate the effectiveness of ABANet, we conducted ablation studies using the proposed dataset as examples.

To better show the influence of the refinement Network and the attention gate, we report the quantitative comparison results in Table 5. First, we conducted the baseline network without any module, which is a segmentation network of our proposed ABANet. Next, we added the refinement network on this baseline. Then, three attention gates are employed in our segmentation network. Finally, attention gate and refinement network

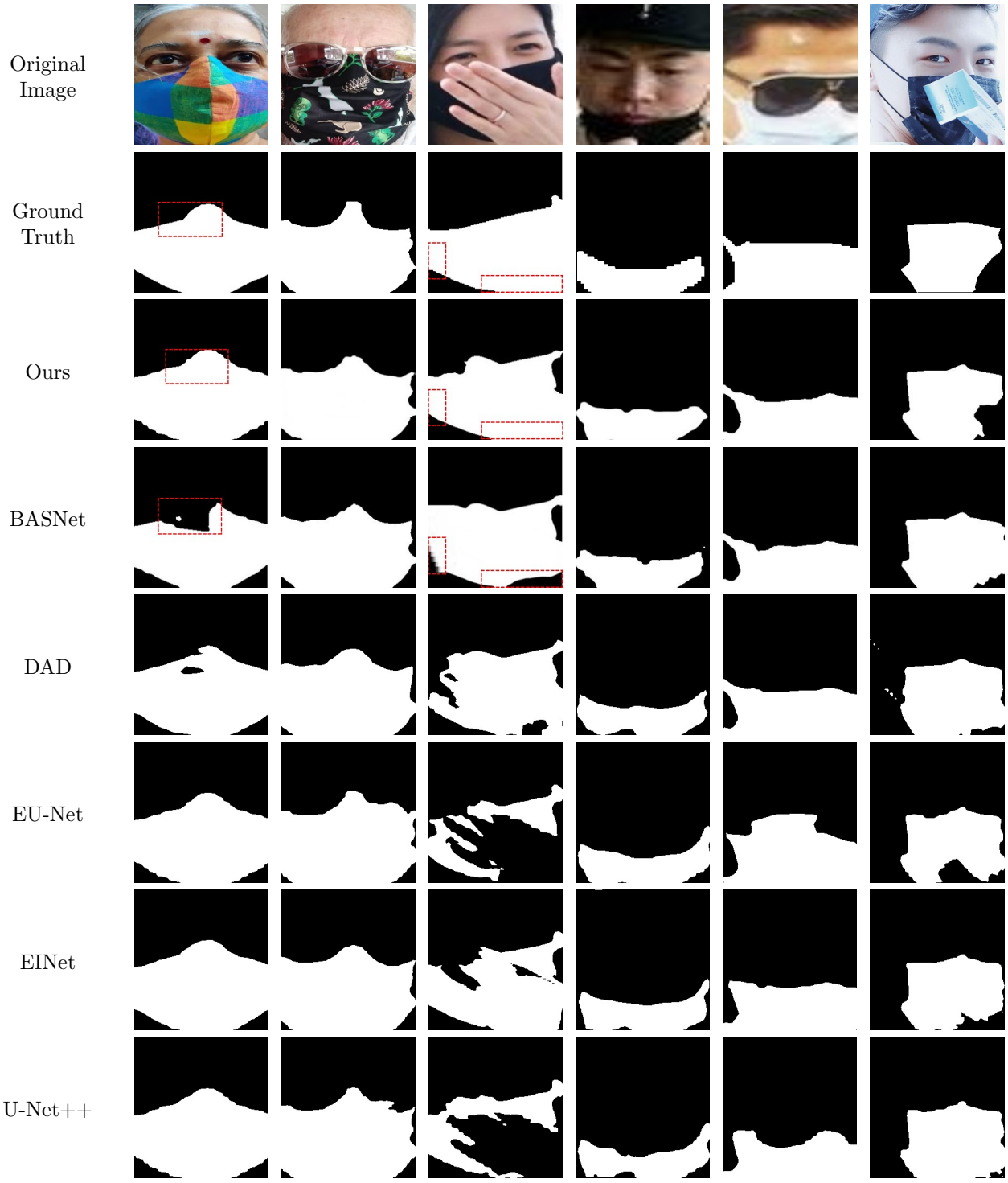


Figure 8: Qualitative comparison of MFSD test set results with cutting-edge image segmentation algorithms. The first two rows display the original photos and the matching ground truth. Rows 3 to 8 illustrate the segmentation results respectively derived from the proposed method, BASNet, DAD, EU-Net, EINet, and U-Net++. Red boxes indicate the fine distinction between the ground truth, the BASNet, and the proposed method.

are combined. Numerical results are shown in Table 5, and the best values are marked in bold. It can be observed that the proposed ABANet brings the most gains in IoU, F1-score, recall, and accuracy. This indicates that AG and the refinement network allow the network to have a stronger ability to detect more changed areas and capture more contextual information respectively.

Table 4: Comparison with masked face segmentation models on our MSFD dataset

Paper	IoU	F1-score	MAE	F_β	FPS
DIN et. al [19]	90.3 ± 0.6	95.7 ± 0.4	0.043 ± 0.02	92.7 ± 0.4	105
Zhang et. al [23]	81.2 ± 1.4	88.66 ± 0.8	0.172 ± 0.08	82.5 ± 0.6	273
Geng et. al [21]	88.9 ± 0.4	94.2 ± 0.3	0.067 ± 0.06	91.9 ± 0.3	115
Sola and Gera [63]	91.4 ± 0.2	96.2 ± 0.3	0.027 ± 0.03	95.6 ± 0.5	66
ours	93.4 ± 0.6	96.5 ± 0.4	0.014 ± 0.02	96.8 ± 0.4	75

Table 5: The performance of different configurations of ABANet.

Method	IoU	F1-score	Precision	Recall	Acc
baseline	91.047	94.99	94.254	95.745	96.185
baseline+refinement	91.73	95.542	94.859	96.235	96.787
baseline+AG	92.425	96.73	**97.28**	96.188	97.256
ABANet	**93.814**	**96.817**	97.164	**96.474**	**97.623**

Table 6: Ablation study results comparing different attention gate jump connection stages in ABANet.

Attention gate jump connection stages				Metrics		
3	4	5	6	IoU	F1-score	Acc
×	×	✓	✓	92.75	95.638	97.308
×	✓	✓	✓	**93.814**	**96.817**	97.623
✓	✓	✓	✓	93.588	96.72	**97.671**

Table 7: ABANet performance with various input picture resolutions.

	Input size		
	128 × 128	256 × 256	512 × 512
IoU	92.576	**93.814**	93.241

Table 8: The performance of different optimizer.

Optimizer	IoU	F1-score	Acc
SGD	93.182	95.72	97.32
Adamax	93.584	96.432	97.531
RMSprop	93.635	96.509	**97.75**
Adam	**93.814**	**96.817**	97.623

Table 9: Effect of various loss functions on ABANet performance.

Losses	IoU	F1-score	Acc
Focal	88.879	92.84	95.4
IoU	90.32	93.79	94.27
Focal + IoU	91.262	94.18	95.9
Focal + SSIM + IoU	**93.814**	**96.817**	**97.623**

In ABANet, attention gates are incorporated to selectively emphasize relevant features while suppressing irrelevant ones, thereby enhancing the segmentation performance. The question addressed in this ablation study concerns the optimal stage at which the attention gate should be introduced within the network architecture. The table 6 showcases the evaluation metrics obtained for different combinations of attention gate involvement across jump connection stages. Each row in the table represents a specific configuration of attention gate engagement, denoted by checkboxes (✓) or crosses (×) corresponding to the presence or absence of the attention gate at each jump connection stage. Overall, the results suggest that integrating attention gates from fourth jump connection stages yields improvements in segmentation performance, as evidenced by higher IoU and F1-score.

For evaluating the stability of the network, we trained the ABANet with 3 different input sizes, including 128 × 128, 256 × 256 and 512 × 512. Based on the data presented in Table 7, the input size of 256 × 256 obtains the best IoU.

In this article, we use Adam as our optimizer. We also comprise different major optimizers: Adam, Adamax, RMSprop, and SGD in Table 8. the Adam optimizer outperformed other optimizers.

In order to illustrate the efficacy of our suggested hybrid loss, we perform an ablation study on losses with the same experimental setup. Table 9 provides the comparing results. It can be observed that both focal loss and IoU loss have similar performance in terms of F1-score and accuracy metrics. However, when it comes to the IoU measure, IoU loss outperforms focal loss. When we mix focal loss and IoU loss, all metrics increased slightly. To propose a final hybrid loss, we use SSIM with two other losses. Table 9 indicates that by equipping hybrid loss on ABANet performance improve greatly. It is due to utilizing SSIM loss and obtaining more detailed boundary prediction.

6. Conclusion

In this article, we suggested an ABANet architecture with a hybrid loss for mask face segmentation. Firstly, we adopted an attention gate in skip connections to capture more specific information. Experimental results show that ABANet can obtain superior segmentation performance by adding a series of AG modules to the skip connections. In addition, the hybrid loss is utilized to evaluate training at the pixel, patch, and map levels, maintaining training stability and adjusting to unbalanced positive and negative sample distributions.

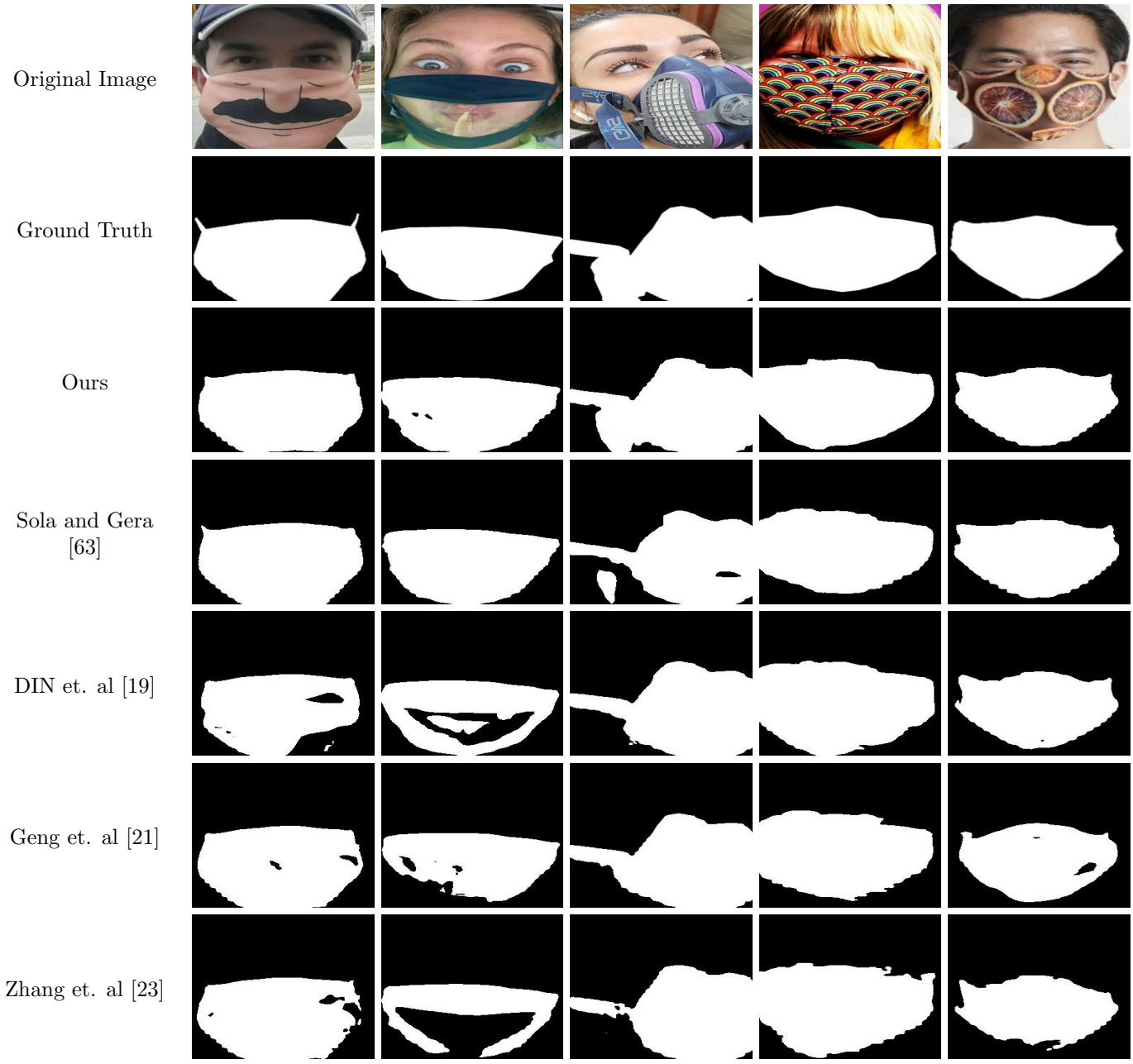


Figure 9: visualization of the inference results obtained by our proposed model under challenging mask designs and shapes, compared to four masked face segmentation models.

There has been minimal progress in masked face-related activities due to the absence of a large-scale, annotated collection of masked faces. To bridge this gap, we proposed a new instance segmentation dataset encompassing 11601 images and 12758 masked faces. We believe this dataset can facilitate the development of face-related tasks with mask occlusion.

To balance network complexity and accuracy gains, we will explore different techniques and improve our attention mechanism in future research. Besides, we intend to design and train a GAN network based on the dataset and method we have proposed to reconstruct mask regions and increase the accuracy of masked face recognition.

References

Y. Sun, X. Wang, X. Tang, Deep learning face representation from predicting 10,000 classes, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 1891–1898.

D.-x. Zhang, P. An, H.-x. Zhang, Application of robust face recognition in video surveillance systems, Optoelectronics Letters 14 (2) (2018) 152–155.

J. S. del Rio, D. Moctezuma, C. Conde, I. M. de Diego, E. Cabello, Automated border control e-gates and facial recognition systems, computers & security 62 (2016) 49–72.

S. R. Jadhav, B. U. Joshi, A. K. Jadhav, Attendance system using face recognition for academic education, in: Computer Networks and Inventive Communication Technologies, Springer, 2021, pp. 431–436.

G. Bargshady, X. Zhou, R. C. Deo, J. Soar, F. Whittaker, H. Wang, Enhanced deep learning algorithm development to detect pain intensity from facial expression images, Expert Systems with Applications 149 (2020) 113305.

S. Koley, H. Roy, S. Dhar, D. Bhattacharjee, Illumination invariant face recognition using fused cross lattice pattern of phase congruency (fclppc), Information Sciences 584 (2022) 633–648.

E. Zangeneh, M. Rahmati, Y. Mohsenzadeh, Low resolution face recognition using a two-branch deep convolutional neural network architecture, Expert Systems with Applications 139 (2020) 112854.

C. Thai, V. Tran, M. Bui, D. Nguyen, H. Ninh, H. Tran, Real-time masked face classification and head pose estimation for rgb facial image via knowledge distillation, Information Sciences 616 (2022) 330–347.

Q. Huang, C. Huang, X. Wang, F. Jiang, Facial expression recognition with grid-wise attention and visual transformer, Information Sciences 580 (2021) 35–54.

D. Zeng, R. Veldhuis, L. Spreeuwers, A survey of face recognition techniques under occlusion, arXiv preprint arXiv:2006.11366.

Y. Long, F. Zhu, L. Shao, J. Han, Face recognition with a small occluded training set using spatial and statistical pooling, Information Sciences 430 (2018) 634–644.

H. Peng, Z. Xing, X. Liu, Z. Gao, H. He, Toward masked face recognition: An effective facial feature extraction and refinement model in multiple scenes, Expert Systems 40 (2) (2023) e13166.

W. Hariri, Efficient masked face recognition method during the covid-19 pandemic, Signal, image and video processing 16 (3) (2022) 605–612.

S. Ge, C. Li, S. Zhao, D. Zeng, Occluded face recognition in the wild by identity-diversity inpainting, IEEE Transactions on Circuits and Systems for Video Technology 30 (10) (2020) 3387–3397. doi:10.1109/TCSVT.2020.2967754.

J. Dong, L. Zhang, H. Zhang, W. Liu, Occlusion-aware gan for face de-occlusion in the wild, in: 2020 IEEE International conference on multimedia and expo (ICME), IEEE, 2020, pp. 1–6.

Y. Li, S. Liu, J. Yang, M.-H. Yang, Generative face completion, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 3911–3919.

P. B. S. Varma, S. Paturu, S. Mishra, B. S. Rao, P. M. Kumar, N. V. Krishna, Sldcnet: Skin lesion detection and classification using full resolution convolutional network-based deep learning cnn with transfer learning, Expert Systems 39 (9) (2022) e12944.

Q. Duan, L. Zhang, Look more into occlusion: Realistic face frontalization and recognition with boostgan, IEEE transactions on neural networks and learning systems 32 (1) (2020) 214–228.

N. U. Din, K. Javed, S. Bae, J. Yi, A novel gan-based network for unmasking of masked face, IEEE Access 8 (2020) 44276–44287.

O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, et al., Attention u-net: Learning where to look for the pancreas, arXiv preprint arXiv:1804.03999.

M. Geng, P. Peng, Y. Huang, Y. Tian, Masked face recognition with generative data augmentation and domain constrained ranking, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 2246–2254.

X. Yin, L. Chen, Non-deterministic face mask removal based on 3d priors, arXiv preprint arXiv:2202.09856.

M. Zhang, R. Liu, D. Deguchi, H. Murase, Masked face recognition with mask transfer and self-attention under the covid-19 pandemic, IEEE Access 10 (2022) 20527–20538.

C. Rother, V. Kolmogorov, A. Blake, " grabcut" interactive foreground extraction using iterated graph cuts, ACM transactions on graphics (TOG) 23 (3) (2004) 309–314.

Y. Zhang, X. Zhang, W. Zhu, Anc: Attention network for covid-19 explainable diagnosis based on convolutional block attention module., CMES-Computer Modeling in Engineering & Sciences 127 (3).

Y.-D. Zhang, Z. Zhang, X. Zhang, S.-H. Wang, Midcan: A multiple input deep convolutional attention network for covid-19 diagnosis based on chest ct and chest x-ray, Pattern recognition letters 150 (2021) 8–16.

A. Fateh, R. T. Birgani, M. Fateh, Unveiling cross-linguistic mastery: Advancing multilingual handwritten numeral recognition with attention-driven transfer learning.

A. Fateh, M. Fateh, V. Abolghasemi, Enhancing optical character recognition: Efficient techniques for document layout analysis and text line detection, Engineering Reports (2023) e12832.

C. Kaul, S. Manandhar, N. Pears, Focusnet: An attention-based fully convolutional network for medical image segmentation, in: 2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019), IEEE, 2019, pp. 455–458.

J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, H. Lu, Dual attention network for scene segmentation, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 3146–3154.

Y. Li, K. Guo, Y. Lu, L. Liu, Cropping and attention based approach for masked face recognition, Applied Intelligence 51 (5) (2021) 3012–3025.

S. Woo, J. Park, J.-Y. Lee, I. S. Kweon, Cbam: Convolutional block attention module, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 3–19.

S. Ge, J. Li, Q. Ye, Z. Luo, Detecting masked faces in the wild with lle-cnns, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2682–2690.

Z. Wang, G. Wang, B. Huang, Z. Xiong, Q. Hong, H. Wu, P. Yi, K. Jiang, N. Wang, Y. Pei, et al., Masked face recognition dataset and application, arXiv preprint arXiv:2003.09093.

G. B. Huang, M. Mattar, T. Berg, E. Learned-Miller, Labeled faces in the wild: A database forstudying face recognition in unconstrained environments, in: Workshop on faces in'Real-Life'Images: detection, alignment, and recognition, 2008.

D. Yi, Z. Lei, S. Liao, S. Z. Li, Learning face representation from scratch, arXiv preprint arXiv:1411.7923.

T. Karras, S. Laine, T. Aila, A style-based generator architecture for generative adversarial networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 4401–4410.

A. Cabani, K. Hammoudi, H. Benhabiles, M. Melkemi, Maskedface-net–a dataset of correctly/incorrectly masked face images in the context of covid-19, Smart Health 19 (2021) 100144.

F. Zhao, J. Feng, J. Zhao, W. Yang, S. Yan, Robust lstm-autoencoders for face de-occlusion in the wild, IEEE Transactions on Image Processing 27 (2) (2017) 778–790.

X. Yuan, I. K. Park, Face de-occlusion using 3d morphable model and generative adversarial network, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 10062–10071.

Y.-J. Ju, G.-H. Lee, J.-H. Hong, S.-W. Lee, Complete face recovery gan: Unsupervised joint face rotation and de-occlusion from a single-view image, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022, pp. 3711–3721.

D.Chiang, Detecting faces and determine whether people are wearing mask (2020).
URL :https://github.com/AIZOOTech/FaceMaskDetection

B. C. Russell, A. Torralba, K. P. Murphy, W. T. Freeman, Labelme: a database and web-based tool for image annotation, International journal of computer vision 77 (1) (2008) 157–173.

X. Qin, D.-P. Fan, C. Huang, C. Diagne, Z. Zhang, A. C. Sant'Anna, A. Suarez, M. Jagersand, L. Shao, Boundary-aware segmentation network for mobile and web applications, arXiv preprint arXiv:2101.04704.

K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

X. Cai, Y. Cao, Y. Ren, Z. Cui, W. Zhang, Multi-objective evolutionary 3d face reconstruction based on improved encoder–decoder network, Information Sciences 581 (2021) 233–248.

Z. Wang, E. P. Simoncelli, A. C. Bovik, Multiscale structural similarity for image quality assessment, in: The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003, Vol. 2, Ieee, 2003, pp. 1398–1402.

O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., Imagenet large scale visual recognition challenge, International journal of computer vision 115 (3) (2015) 211–252.

P. Yakubovskiy, Segmentation models pytorch (2020).
URL https://github.com/qubvel/segmentation_models.pytorch

D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980.

A. Fateh, M. Fateh, V. Abolghasemi, Multilingual handwritten numeral recognition using a robust deep network joint with transfer learning, Information Sciences 581 (2021) 479–494.

D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, A. Borji, Structure-measure: A new way to evaluate foreground maps, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 4548–4557.

Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, P. H. Torr, Deeply supervised salient object detection with short connections, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 3203–3212.

O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, Springer, 2015, pp. 234–241.

Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, J. Liang, Unet++: A nested u-net architecture for medical image segmentation, in: Deep learning in medical image analysis and multimodal learning for clinical decision support, Springer, 2018, pp. 3–11.

A. Chaurasia, E. Culurciello, Linknet: Exploiting encoder representations for efficient semantic segmentation, in: 2017 IEEE Visual Communications and Image Processing (VCIP), IEEE, 2017, pp. 1–4.

T. Fan, G. Wang, Y. Li, H. Wang, Ma-net: A multi-scale attention network for liver and tumor segmentation, IEEE Access 8 (2020) 179656–179665.

H. Li, P. Xiong, J. An, L. Wang, Pyramid attention network for semantic segmentation, arXiv preprint arXiv:1805.10180.

C. Li, G. Jiao, Einet: camouflaged object detection with pyramid vision transformer, Journal of Electronic Imaging 31 (5) (2022) 053002.

K. Patel, A. M. Bur, G. Wang, Enhanced u-net: A feature enhancement network for polyp segmentation, in: 2021 18th Conference on Robots and Vision (CRV), IEEE, 2021, pp. 181–188.

J. Li, W. He, H. Zhang, Towards complex backgrounds: A unified difference-aware decoder for binary segmentation, arXiv preprint arXiv:2210.15156.

L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 801–818.

S. Sola, D. Gera, Unmasking your expression: Expression-conditioned gan for masked face inpainting, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 5907–5915.

K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask r-cnn, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 2961–2969.