Heliyon 10 (2024) e37338

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

GU-Net: Causal relationship-based generative medical image segmentation model

Dapeng Cheng [a,b,*], Jiale Gai [a], Bo Yang [e], Yanyan Mao [a], Xiaolian Gao [a], Baosheng Zhang [a], Wanting Jing [c], Jia Deng [a], Feng Zhao [a,b], Ning Mao [d]

[a] School of Computer Science and Technology, Shandong Technology and Business University, Yantai, 264005, Shandong, China
[b] Shandong Co-Innovation Center of Future Intelligent Computing, Yantai, 264005, Shandong, China
[c] School of Information and Electronic Engineering, Shandong Business and Technology University, Yantai, 264005, Shandong, China
[d] Department of Radiology, Yantai Yuhuangding Hospital, Yantai, 264000, Shandong, China
[e] Hainan College of Economics and Business School of Information Technology, Haikou, 571127, Hainan, China

ARTICLE INFO

Keywords:
Medical image segmentation
Convolutional neural network
Attention mechanism
Causal reasoning
Interactive training

ABSTRACT

Due to significant anatomical variations in medical images across different cases, medical image segmentation is a highly challenging task. Convolutional neural networks have shown faster and more accurate performance in medical image segmentation. However, existing networks for medical image segmentation mostly rely on independent training of the model using data samples and loss functions, lacking interactive training and feedback mechanisms. This leads to a relatively singular training approach for the models, and furthermore, some networks can only perform segmentation for specific diseases. In this paper, we propose a causal relationship-based generative medical image segmentation model named GU-Net. We integrate a counterfactual attention mechanism combined with CBAM into the decoder of U-Net as a generative network, and then combine it with a GAN network where the discriminator is used for backpropagation. This enables alternate optimization and training between the generative network and discriminator, enhancing the expressive and learning capabilities of the network model to output prediction segmentation results closer to the ground truth. Additionally, the interaction and transmission of information help the network model capture richer feature representations, extract more accurate features, reduce overfitting, and improve model stability and robustness through feedback mechanisms. Experimental results demonstrate that our proposed GU-Net network achieves better segmentation performance not only in cases with abundant data samples and relatively simple segmentation targets or high contrast between the target and background regions but also in scenarios with limited data samples and challenging segmentation tasks. Comparing with existing U-Net networks with attention mechanisms, GU-Net consistently improves Dice scores by 1.19%, 2.93%, 5.01%, and 5.50% on ISIC 2016, ISIC 2017, ISIC 2018, and Gland Segmentation datasets, respectively.

* Corresponding author at: School of Computer Science and Technology, Shandong Technology and Business University, Yantai, 264005, Shandong, China.
E-mail addresses: chengdapeng@sdtbu.edu.cn (D. Cheng), gjl0108@163.com (J. Gai), yangbo_711@163.com (B. Yang), maoyanyan@sdtbu.edu.cn (Y. Mao), 15610665770@163.com (X. Gao), 1307084694@qq.com (B. Zhang), 2804727713@qq.com (W. Jing), 2907845396@qq.com (J. Deng), zhaofeng1016@126.com (F. Zhao), maoning@pku.edu.cn (N. Mao).

https://doi.org/10.1016/j.heliyon.2024.e37338
Received 16 September 2023; Received in revised form 31 August 2024; Accepted 2 September 2024

Available online 4 September 2024

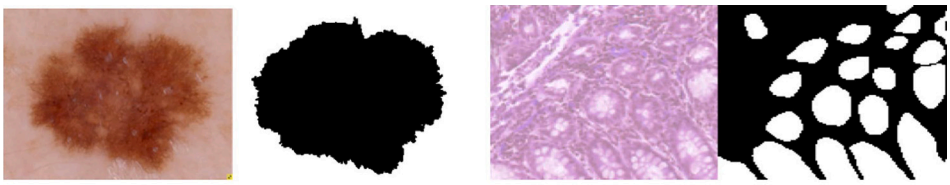

Fig. 1. Original images and corresponding ground-truth annotations in medical image segmentation tasks. (a) shows the original images of skin lesions along with their corresponding ground-truth images. (b) illustrates the tissue slice images for glandular segmentation along with their corresponding ground-truth images.

1. Introduction

Medical images play a crucial role in the diagnosis and treatment of diseases. The emergence of computer-aided diagnosis and treatment systems has provided more accurate medical image interpretation for various diseases, assisting doctors in more effective treatment [1–3]. Segmenting organs or lesion images from medical scans aids doctors in making accurate judgments, planning surgical procedures, and devising appropriate treatment strategies. Moreover, precise and reliable medical image processing reduces the time, cost, and errors associated with manual processing. Among numerous medical image processing tasks, medical image segmentation is a critical component [4]. While medical image segmentation helps to clarify changes in anatomical or pathological structures in images, aiding physicians in identifying missed lesions and improving disease prevention and treatment plans, it remains a challenging task due to the complex geometric structures, blurred edges, various types of interference, and inherent noise values present in medical images.

As shown in Fig. 1, medical image segmentation has been applied in the fields of skin lesion segmentation [5] and glandular segmentation [6]. In (a), it depicts original skin lesion images and their corresponding ground-truth images, while in (b), it illustrates tissue slice images for glandular segmentation along with their corresponding ground-truth images. When performing segmentation of skin lesion images, a number of difficulties still need to be overcome, even though sufficient data samples and a high contrast between the lesion area and the background in some images can be utilized. These difficulties include the fact that the size of the lesion area may be small, the edges may not be sufficiently sharp, and the presence of interfering factors such as hairs and discoloration on the skin, all of which can make the segmentation process more complex. For the segmentation of the glandular dataset, the small number of samples, the extremely low contrast between the glandular region and the background, the unclear boundaries between the glands, and the possible presence of structures within the glands that are similar to those in the background combine to make the segmentation process more difficult.

Deep learning networks have made remarkable achievements in the field of medical image segmentation, surpassing state-of-the-art non-deep learning methods. Fully Convolutional Networks (FCNs) [7] were among the earliest deep learning networks applied in medical image segmentation. Ronneberger et al. extended this architecture and proposed the U-Net [8] network structure, which achieved good segmentation performance with a large amount of training data. They consist of an encoder and a decoder. In the encoding stage, image features are primarily extracted through downsampling. In the decoding stage, upsampling is performed to generate segmentation maps of the same size as the input. Currently, various networks such as V-Net [9], 3D U-Net [10], Res-UNet [11], Dense-UNet [12], YNet [13], KiU-Net [14,15], and U-Net3+ [16] have been specifically proposed for image and volume segmentation of various medical imaging modalities. The utilization of attention mechanisms in U-Net models [17,18] enables the networks to focus more on the most relevant information in the feature maps while suppressing irrelevant information. These techniques have demonstrated excellent performance on many challenging datasets, confirming the efficiency of convolutional neural networks in recognizing and extracting organ or lesion features from medical scans. However, these network models rely on a huge amount of data to be trained and a large number of network parameter configurations to exhibit good generalization capabilities. This leads to the fact that current segmentation methods are largely dependent on data samples and are sensitive to the information in the input data. However, an important challenge faced in the field of medical image segmentation is the difficulty of obtaining sufficient data samples. In addition, most of the current attention mechanisms applied to medical image segmentation are based on traditional probability-based visual learning methods, which are only used to explicitly guide the final prediction results, while ignoring the causal connection between the prediction and the attention mechanism. Specifically, existing attention mechanisms mainly perform weakly supervised training via loss functions and lack other supervisory signals to guide the training process. Such likelihood-based methods focus only on explicit guidance of the final prediction without considering the causal interaction between prediction and attention.

We propose the GU-Net (Generative U-Net Network) model, where an improved U-Net network serves as the generative network. Specifically, we incorporate the Causal Attention Mechanism based on Channel and Spatial Dimensions (CSCA), which is a fusion of the Convolutional Block Attention Module (CBAM), into the decoder stage. CBAM enables the extraction of image features in both channel and spatial dimensions, emphasizing more relevant features while suppressing less relevant ones. By combining CBAM with the causal attention mechanism, the attention module is not solely dependent on the loss function for training but can also adjust based on the causal relationship between prediction and attention. Additionally, the integration with the discriminator in the Generative Adversarial Networks (GAN) allows the GU-Net to leverage backpropagation. This enables the network to update its parameters without direct reliance on data samples, thus improving the network's generalization performance. Furthermore, the interaction between the two networks in the optimization process enhances the accuracy of segmentation results. We will conduct

medical image segmentation experiments on both the skin lesion image dataset and the GlaS gland dataset. These two datasets are vastly different in type, with varying sizes and shapes of areas to be segmented, and they exhibit significant interference factors. This will validate that GU-Net can achieve good segmentation results not only with sufficient data but also with smaller datasets. Additionally, conducting experiments on different types of datasets will verify the generalization performance of GU-Net.

2. Related work

2.1. Medical image segmentation

Early methods for medical image segmentation typically relied on edge detection, template matching, statistical shape models, active contours, and traditional machine learning techniques. These methods have achieved decent results to some extent, but medical images often suffer from issues such as blurriness, noise, low contrast, making feature representation more difficult. Therefore, medical image segmentation remains one of the most challenging tasks in the field of computer vision.

Due to the rapid advancement of deep learning techniques, manual feature extraction for medical image segmentation is no longer widely employed. Convolutional neural networks (CNN) have successfully achieved hierarchical feature representation of images, making them one of the hottest research topics in image processing and computer vision. Since CNNs, which are used for feature learning, are not sensitive to image noise, blurriness, and contrast, they can achieve excellent results in medical image segmentation.

Neural networks used for image segmentation typically adopt the encoder-decoder architecture of fully convolutional networks. The encoder is responsible for extracting image features, while the decoder aims to restore the extracted features to the original image size and output the final segmentation result. Typical neural network structures include U-Net [8], FCN [7], Deeplab [19], among others. U-Net is an end-to-end trainable network composed of an encoder and a decoder. The encoder functions as a feature extractor, acquiring high-dimensional features across multiple scales, which the decoder utilizes to reconstruct the target for segmentation. U-Net incorporates skip connections to effectively fuse low-resolution and high-resolution feature maps, effectively combining image features at different resolutions. Currently, U-Net has become a benchmark model for most medical image segmentation tasks and has undergone various efficient improvements. For instance, Zhou et al. proposed the U-Net++ model [20]. Compared to U-Net, U-Net++ combines the DenseNet structure, enhancing gradient flow through dense skip connections. Additionally, it introduces multi-level upsampling and concatenation operations to increase network capacity and contextual information, thereby improving segmentation performance on medical images. Milletari et al. introduced the V-Net model [9], which achieves 3D medical image segmentation. Compared to U-Net, V-Net utilizes 3D convolutional kernels to better exploit the high-dimensional spatial correlation in the data. Furthermore, V-Net employs residual connections for four layers of convolutional operations, resulting in better performance in medical image segmentation. AttResDU-Net [21] proposes an attention-based residual dual U-Net architecture, which modifies existing medical image segmentation models by incorporating attention gates on skip connections and residual connections within the convolutional blocks. However, existing neural networks are trained based on input data and are heavily influenced by it. If the model can be combined with a GAN network's discriminator, it can utilize the discriminator for backpropagation, updating only using gradients flowing through the discriminator, thereby enhancing the model's generalization ability.

In recent years, the emergence of Transformer network models has led to various adaptations in medical image segmentation models. For instance, U-Net Transformer [22] applies the U-shaped architecture for image segmentation while integrating Transformer's self-attention and cross-attention mechanisms. Although visual Transformers have made significant progress in the field of image segmentation, it's important to recognize that most visual Transformer models are relatively complex, computationally intensive, and require longer training times, demanding higher hardware specifications. In contrast, our proposed GU-Net model is a lightweight network, requiring lower hardware specifications, shorter training times, and is more suitable for real-time applications.

2.2. Attention mechanisms

Attention is one of the fundamental mechanisms in human visual perception. When confronted with complex scenes, humans have the ability to selectively focus on regions of interest, narrowing down the search range and accelerating recognition speed. Many researchers have made efforts to simulate human attention mechanisms in computer vision systems [23–25], aiming to facilitate more accurate recognition by discovering regions of interest and alleviating negative effects caused by variations in visual appearance, cluttered backgrounds, occlusions, and so on.

Existing attention mechanisms can be broadly categorized into two types: local attention and non-local attention. Oktay et al. proposed an attention U-Net model [18], which incorporates a spatial attention block into the U-Net architecture. By using spatial attention, weight maps are generated to modulate the output of the encoder, suppressing low-correlation regions and highlighting salient features for pancreas segmentation. SA-UNet [26] achieves state-of-the-art performance by introducing spatial attention modules, inferring attention maps along the spatial dimension, and multiplying them with input feature maps to achieve adaptive feature refinement. This has been demonstrated on the DRIVE and CHASE_DB1 datasets for vessel segmentation. Channel attention, on the other hand, leverages learned global information to selectively emphasize useful features and suppress irrelevant ones, adjusting the response strength of features across channels. Hu et al. introduced Squeeze-and-Excitation Networks (SE-Net) [27], which employ a three-step process involving squeezing, exciting, and generating feature weight maps using the sigmoid function to achieve channel-wise attention weighting, revealing the importance of channel features across the entire feature map. While spatial attention overlooks the variations in different channels and treats each channel equally, channel attention directly focuses on global information, disregarding spatial local information for each channel. Therefore, researchers have proposed models with hybrid attention

blocks, combining the advantages of both attention mechanisms. For instance, Woo et al. proposed the Convolutional Block Attention Module (CBAM) [28], which combines spatial attention and channel attention modules. Compared to attention mechanisms that solely focus on either spatial or channel aspects, CBAM achieves better results. Non-local attention mechanisms distribute weights across different parts of the entire input signal, allowing the model to concentrate on crucial parts to improve performance. Wang et al. introduced a non-local U-Net to overcome the limitations of convolutional operations in medical image segmentation [29]. This network model utilizes self-attention mechanisms and global aggregation blocks during the upsampling and downsampling processes to extract comprehensive image information, leading to more accurate segmentation predictions. Non-local attention mechanisms capture contextual information of input signals more comprehensively compared to local attention mechanisms, which only focus on a subset of the data. However, local attention mechanisms are generally more efficient as they only need to attend to specific regions in the data, in contrast to non-local attention mechanisms.

It can be seen that attention mechanisms are effective in improving the segmentation accuracy of medical images. However, most attention modules currently used in medical image segmentation are supervised by the final loss function, lacking strong supervision signals to guide the training process. They are only used to explicitly train the generation of the final prediction, disregarding the causal relationship between prediction and attention mechanism. In GU-Net, a generative network based on the U-Net model is employed as the underlying framework, and a causal counterfactual attention mechanism fused with the CBAM attention module is introduced in the decoding stage for medical image segmentation. The counterfactual attention mechanism [30] quantifies the quality of attention by comparing the impact of facts (features highly correlated with the ground-truth) and counterfactuals (features with lower correlation with the ground-truth) on the final prediction. Then, by maximizing the differences, the counterfactual attention mechanism enables the network to learn the most effective visual features, reducing the influence of interference factors on medical image segmentation and enhancing the discriminative power of the network model.

3. Methods

Taking inspiration from interactive training and the causal relationship between attention mechanisms, we propose the GU-Net network model as shown in Fig. 2. This model combines the advantages of spatial and channel attention mechanisms for feature extraction, the causal relationship between prediction and attention, and the interactive training between the generator and discriminator in GAN networks. In the following sections, we will provide a detailed description of the different components of this network.

3.1. Overall framework of GU-Net

The structure of the GU-Net network model is illustrated in Fig. 2. While maintaining generality, we chose the powerful U-Net network model [8] as the base model for the generator in our approach, integrating four Counterfactual Spatial Channel Attention (CSCA) mechanisms fused with CBAM into the decoder part of the U-Net. This allows each attention module to correspond to a level or a feature map within the decoder. The main advantages are as follows: (1) Each level of the U-Net decoder contains features at different scales. By incorporating attention mechanisms at each level, the network can better focus on information at specific scales, enabling the fusion of multi-scale features. (2) Applying attention mechanisms at each level enhances the network's ability to accurately reconstruct details and boundary information of the target objects. Each level's attention mechanism helps the network concentrate on processing information at specific scales, thereby improving its reconstruction capability. (3) Attention mechanisms aid in extracting relevant feature information while suppressing irrelevant details. Introducing attention mechanisms at each level enhances the network's representational capacity, enabling more effective learning of image representations. (4) Feature maps at different levels may have varying levels of information and importance. Integrating attention mechanisms at each level allows the network to adaptively adjust attention allocation based on the feature information at different levels, thereby better adapting to different types of input images. In summary, adding four attention modules allows the generator to flexibly and effectively handle features at different scales, thereby improving segmentation performance and enhancing the network's generalization capability. The GU-Net model is further combined with a discriminator to achieve interactive training through backpropagation. The input to the generator network is either the original lesion image or tissue slice image, while the input to the discriminator consists of the concatenation of the original lesion image or tissue slice image with the generated result from the generator network, and the concatenation of the original lesion image or tissue slice image with its corresponding ground truth image. In the GU-Net model, the improved U-Net network as the generator aims to generate segmentation result maps with high accuracy to deceive the discriminator network, while the objective of the discriminator is to distinguish between the generated feature maps by the generator and the ground-truth. In this way, the generator and discriminator form a dynamic "game process". During the "game process", the generator and discriminator continuously update their weights to generate feature maps that are most correlated with the ground-truth.

3.2. CSCA attention mechanism

The purpose of the CSCA module is to enhance the quality of the model by adjusting the prediction based on the causal relationship between prediction and attention, building upon the accurate segmentation feature maps extracted by CBAM in both channel and spatial dimensions, as illustrated in Fig. 3.

The CBAM (Convolutional Block Attention Module) effectively captures global contextual information and channel-wise correlations of image features. It consists of two sub-modules: the channel attention module and the spatial attention module.

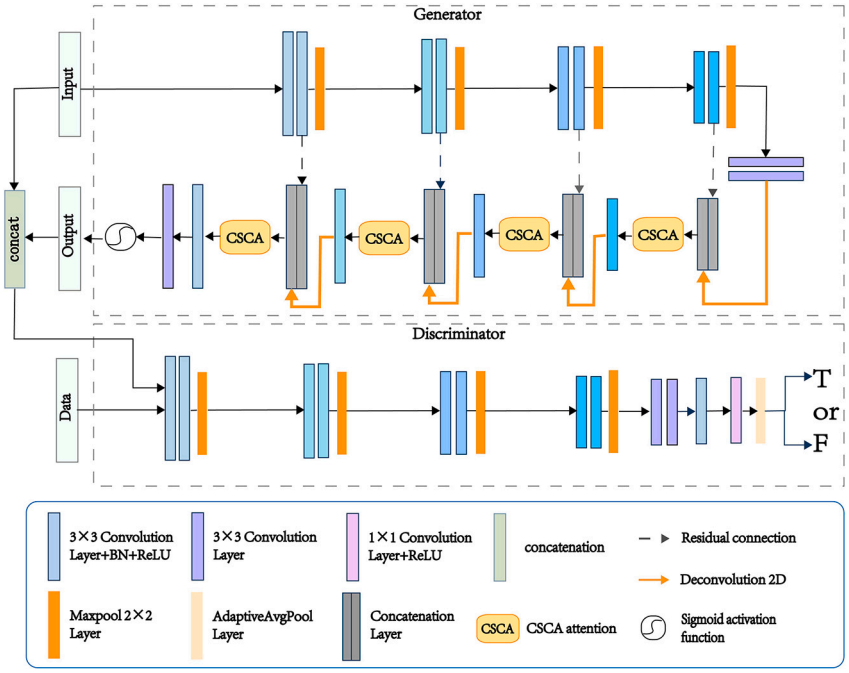

Fig. 2. Overall architecture of GU-Net.

The channel attention module aims to adaptively learn the importance of each channel and enhance relevant feature channels. It achieves this by modeling global information for each channel. Initially, the channel attention module employs adaptive global max pooling and adaptive global average pooling operations to extract the maximum and average values for each channel from the input feature map. Subsequently, these globally pooled features are passed through a multilayer perceptron (MLP) to learn the importance weights for each channel. The MLP typically comprises two convolutional layers followed by a ReLU activation function. The first convolutional layer reduces the number of input channels to 1/16 of the original size, followed by ReLU activation, and the final convolutional layer restores the number of channels to the original size to learn the weights for each channel. The output values are then compressed to a range between 0 and 1 using the sigmoid function, representing the attention weights for each channel. Finally, the channel attention weights are applied to each channel of the input feature map to enhance useful feature channels through weighted multiplication.

The feature map processed by the channel attention mechanism serves as the input to the spatial attention module. The spatial attention module aims to adaptively learn the importance of each spatial position and enhance relevant spatial regions. It achieves this by modeling features for each position. Initially, the maximum and average values of the input feature map along the channel dimension are computed separately. These results are concatenated along the channel dimension using the concat operation to form a feature map with two channels. Then, a 7×7 convolution operation is applied to the concatenated feature map to learn the importance of each spatial position, followed by the application of a sigmoid activation function to limit the output between 0 and 1, representing the importance weights for each spatial position. Finally, the spatial importance weights are multiplied with the input feature map to obtain the combined result after channel and spatial attention.

When convolutional neural networks utilize the CBAM attention mechanism to improve segmentation accuracy, the optimization of the attention module is solely based on the final prediction results, treating the model as a "black box" and neglecting how the learned attention map influences the prediction results. Causal reasoning helps us break out of the "black box" and analyze the causal relationship between prediction and attention.

A causal graph, also known as a structural causal model, can be used to represent the variables of an attention model, including the input feature map X, the attention-generated feature map A, and the final prediction F. The connection $X\rightarrow A$ indicates that the feature map X is used as the input for attention, resulting in the corresponding attention feature map A. $(X,A)\rightarrow F$ represents that both the feature map and attention map together determine the final prediction F. Node X is the causal parent of A, and node Y is the causal child of both X and A. With a causal graph, we can perform "intervention" operations, which involve directly manipulating the values of certain variables to analyze causal relationships and observe the resulting effects, denoted as "$do(\bullet)$". When studying the influence of a variable, intervention can be performed by clearing all incoming links to the variable and assigning a specific value to it. For example, in the causal graph, $do(A=\bar{A})$ means requiring variable A to take the value $\bar{A}$ and severing the link $X\rightarrow A$, thereby forcing the variable A to no longer be influenced by its causal parent X. In the CSCA attention mechanism, counterfactual interventions $do(A=\bar{A})$ are performed by generating a non-existing attention map using random numbers while keeping the input feature map unchanged, as illustrated in Fig. 4. The actual impact of learned attention on the prediction can be represented by the

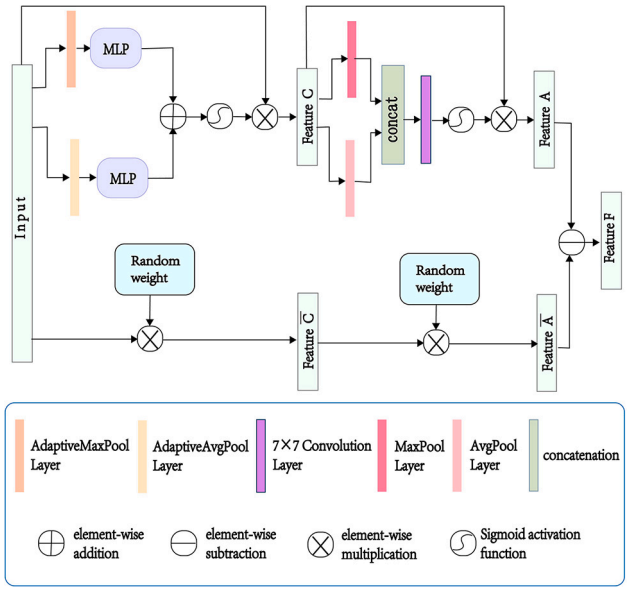

Fig. 3. Structure of the CSCA attention mechanism.

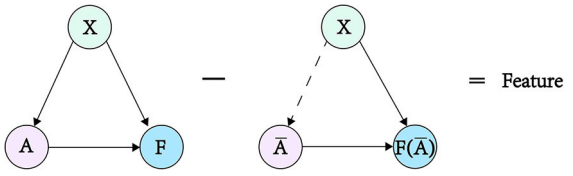

Fig. 4. Illustration of the "intervention" operation in the causal graph.

difference between the observed prediction $Y(A=A,X=X)$ and its counterfactual substitute $Y(do(A=\bar{A}),X=X)$[31–33], as expressed in Equation (1):

$$Feature = E_{\bar{A}\sim\gamma}[Y(A=A,X=X) - Y(do(A=\bar{A}),X=X)] \tag{1}$$

Here, γ represents the distribution of counterfactual attention mechanism. Therefore, the effectiveness of attention can be explained as the difference between effective attention and spurious attention, leading to an improvement in the final prediction.

Thus, the feature map obtained through the CSCA attention mechanism can be represented by Formula (2):

$$F = A - \bar{A} \tag{2}$$

Here, F represents the feature map processed by the CSCA attention module, A represents the feature map processed by the CBAM attention module, and $\bar{A}$ represents the incorrect feature map. A undergoes weighted processing through channel attention and spatial attention, enhancing the importance of useful information contained within it. On the other hand, $\bar{A}$ may contain some unnecessary or irrelevant information due to improper application of channel and spatial weights. Therefore, we utilize the principle of 'intervention' in causal relationships to subtract A from $\bar{A}$, thereby removing features with lower relevance. Specifically, the erroneous attention map $\bar{A}$ is generated using random weights. During training, the generated attention weights are uniformly distributed random numbers between 0 and 2. Then, these randomly initialized attention weights are multiplied with the input features to obtain the feature map $\bar{A}$. If the model is in a non-training state, all generated attention weights are set to 1. Then, the input features are multiplied by these tensors of all 1s to obtain the feature map $\bar{A}$.

By optimizing the attention mechanism, we aim to improve predictions based on erroneous attention and encourage the attention model to discover the most discriminative regions, thereby enhancing the segmentation performance of the generator.

3.3. Loss function

The Dice loss function [9] is commonly used in image segmentation tasks, especially in cases of class imbalance. The Dice loss measures the similarity between the predicted segmentation result and the ground truth segmentation mask based on the overlap region. It is calculated based on the Dice coefficient. During the training of the generative network, the Dice coefficient is designed to measure the overlap between the generated segmentation result and the ground truth, with values ranging from 0 to 1, where 1 indicates perfect overlap and 0 indicates no overlap. To convert the Dice coefficient into a loss function, the Dice loss, as shown

in Formula (3), is typically used. As the Dice coefficient increases, the Dice loss decreases, thereby driving network optimization to improve the accuracy of the segmentation results. Because the Dice loss function does not depend on the absolute values of the predicted results, it performs well for small targets and can better handle class imbalance situations.

The adversarial loss function [34] is commonly used in Generative Adversarial Networks to train the adversarial process between the generator and the discriminator. The generator aims to produce realistic samples, while the discriminator aims to accurately distinguish between generated samples and real samples. The adversarial loss function works by minimizing the adversarial difference between the generator and the discriminator, thereby encouraging the generator to produce more realistic samples while making it difficult for the discriminator to differentiate between generated and real data. Through this adversarial training process, the generator and discriminator engage in a dynamic game. The generator strives to produce more realistic samples to deceive the discriminator, while the discriminator endeavors to improve its ability to discriminate between real and generated samples. Ultimately, this game process leads the generator to produce samples that are more similar to real data, thereby enhancing the performance of the generative network.

In summary, the selection of Dice loss function and adversarial loss function aims to optimize the performance of medical image segmentation tasks. The Dice loss function handles class imbalance and segmentation of small objects, while the adversarial loss function promotes the generation of predicted segmentation results that are closer to real samples, thereby improving the segmentation accuracy of the generative network. The following sections will provide detailed explanations of these two loss functions.

(1) Loss in the Generative U-Net network: Dice loss function. It is represented by Formula (3):

$$Dice\ Loss = 1 - Dice = 1 - \frac{2\left|X \bigcap Y\right|}{|X| + |Y|} \tag{3}$$

Here, X represents the set of pixels in the ground-truth, Y represents the set of pixels in the predicted segmentation image, $\left|X \bigcap Y\right|$ approximates the dot product between the pixels of the ground-truth and the predicted image, and the dot product results are summed. $|X|$ and $|Y|$ approximate the pixel-wise summation in their respective images.

(2) Loss function generated by the adversarial process: Adversarial loss. It is represented by Formula (4):

$$\min_G \max_D V(D,G) = E_{x\sim p_{data}(x)}[\log D(x)] + E_{z\sim p_z(z)}[\log(1 - D(G(z)))] \tag{4}$$

Where x is a sample drawn from the real data distribution p_{data}, z is a sample drawn from the noise distribution p_z, and $G(z)$ represents the fake samples generated by the generator. We train the discriminator to maximize the probability of assigning correct labels to training examples and generator's samples, while training the generator to minimize $\log(1 - D(G(z)))$. In other words, the discriminator and the generator engage in a minimax game, with the value function $V(G, D)$. Ultimately, our goal is to make the generated data distribution match the real distribution.

4. Experimental results

To validate the accuracy and generalization of the GU-Net network model in medical image segmentation, we conducted experiments on three skin lesion image datasets (ISIC 2016, ISIC 2017, ISIC 2018) and the GlaS gland dataset. The experiments included ablation studies of GU-Net on the four datasets and comparative experiments with four related methods.

4.1. Experimental evaluation and methods

The GU-Net experimental model was implemented in the PyTorch framework. We used the Adaptive Moment Estimation (Adam) with an initial learning rate of 10^{-4} for training, a weight decay of 10^{-8}, batch size of 8, and trained for 300 epochs. The learning rate was decayed by 0.5 every 256 steps. The feature channel number in the first block of GU-Net was set to 16 and doubled after each downsampling. The training was conducted on an NVIDIA GEFORCE RTX 3060 GPU. We trained the network using Dice loss function and adversarial loss function, and tested the best-performing model on the validation set at all stages. We performed 5-fold cross-validation for final evaluation.

The quantitative evaluation parameters for segmentation accuracy are: (1) The similarity score between predicted values and ground-truth (Dice); (2) Average symmetric surface distance (ASSD); (3) The accuracy of the positional information of the predicted results. (IoU); (4) Actual inference time of the neural network model (inference time).

4.2. Skin lesion image segmentation

4.2.1. Dataset

We used publicly available datasets of skin lesion images, namely ISIC 2016 [35], ISIC 2017 [36], and ISIC 2018 [37,38]. The ISIC 2016 dataset consists of 900 training images and 379 validation images. The ISIC 2017 dataset includes 2000 training images, 150 validation images, and 600 test images. The ISIC 2018 dataset contains 2594 training images, which we randomly split into 1816, 260, and 518 for training, validation, and testing, respectively. For each sample, the original image and its corresponding ground-truth annotation (including cancerous or non-cancerous lesions) are available. Due to the varying sizes of images in the dataset, we resized each image to 256×342 and normalized them using mean and standard deviation. During training, we performed random cropping of size 224×300, horizontal and vertical flipping, and random rotation within a certain angle range $(-\pi/6, \pi/6)$ for data augmentation.

Table 1
Quantitative Comparison of Network Models with Different Modules and Overall Network in ISIC Skin Lesion Image Segmentation Task.

Datasets	Network	Dice(%)↑	ASSD(pix)↓	IoU(%)↑	Inference time(s)↓
ISIC2016	**Baseline**	92.75	0.5028	87.29	**2.2051**
	Discriminator	92.91	0.6418	87.30	2.2886
	CSCA	93.54	0.7189	88.63	3.0789
	GU – Net	**93.94**	**0.4283**	**88.98**	3.3289
ISIC2017	**Baseline**	86.82	1.6888	78.38	3.5597
	Discriminator	87.56	1.5175	79.33	3.6695
	CSCA	89.30	1.2458	81.76	**3.3925**
	GU – Net	**89.75**	**1.1525**	**82.41**	5.9833
ISIC2018	**Baseline**	87.41	1.0887	78.73	**3.0948**
	Discriminator	92.18	0.6774	86.02	3.9521
	CSCA	92.15	0.7399	86.05	3.6675
	GU – Net	**92.42**	**0.7057**	**86.46**	3.7532

4.2.2. Experimental setup for ablation study

In this section, we perform quantitative and visual comparisons between GU-Net and network models that only include the CSCA attention mechanism, as well as those that only include a discriminator, on the ISIC skin lesion image dataset to investigate the effectiveness of the two modules. Here, CSCA refers to a network model that employs only counterfactual attention mechanisms based on channel and spatial dimensions, while the discriminator represents a network model that uses a discriminator exclusively in interactive training.

The quantitative comparison between the GU-Net network and network models that incorporate either only the CSCA attention mechanism or only the discriminator is shown in Table 1. The models are evaluated on the task of ISIC skin lesion image segmentation. As evidenced in Table 1, the incorporation of both the CSCA and the discriminator resulted in enhanced performance when compared to the U-Net baseline model. Specifically, GU-Net performs exceptionally well in terms of Dice score and IoU, while demonstrating lower ASSD values compared to other variants. On the ISIC 2016 dataset, GU-Net achieves a Dice score of 93.94% and an IoU of 88.98%, with an ASSD value of 0.4283 pix. On the ISIC 2017 dataset, GU-Net achieves a Dice score of 89.75% and an IoU of 82.41%, with an ASSD value of 1.1525 pix. On the ISIC 2018 dataset, GU-Net achieves a Dice score of 92.42% and an IoU of 86.46%, with an ASSD value of 0.7057 pix. These results validate the effectiveness of our proposed method. In Table 1, we have added arrows indicating that higher/lower values are preferable, and bolded the best results. However, due to the interactive training between the two networks inherent to GU-Net, the inference time is longer than that of network models that incorporate only the CSCA attention mechanism or only include the discriminator.

Fig. 5 presents the visual results of our proposed GU-Net network, along with the network models that only have the CSCA attention module and only have the discriminator, in ISIC skin lesion image segmentation. Each image in the figure includes the original image from the ISIC skin lesion image dataset (origin), the corresponding ground-truth, and the predicted segmentation results from different network structures. We selected three representative sets of images from the experimental results on the ISIC 2016, ISIC 2017, and ISIC 2018 datasets. These sets represent three different scenarios: images with edge blurring, other pigment interference or markings, and significant hair interference in the mirror images. The lesion regions to be segmented are outlined in green boxes. The first three images represent the segmentation results for the scenarios of blurred lesion edges, hair interference, and irregular lesion regions with high demand for detailed segmentation, respectively, obtained from the ISIC 2016 dataset. The next three images represent the segmentation results for the scenarios of relatively small and blurred lesion regions, low contrast between the lesion and the patient's skin with interference from skin color patches, and artificially marked lesion regions, respectively, obtained from the ISIC 2017 dataset. The last three images represent the segmentation results for the scenarios of large but blurred lesion regions with low contrast regions within the lesion, small lesion regions with markings and color patches interference, and high hair interference within the segmentation regions, respectively, obtained from the ISIC 2018 dataset.

Through observation of three sets of visualized results, it is evident that the GU-Net model demonstrates more precise handling capabilities in the segmentation prediction task of ISIC skin lesion images compared to other models. Specifically, the GU-Net model exhibits more accurate performance in handling lesion edges and detailed areas, presenting more refined segmentation results. Additionally, the GU-Net model demonstrates strong resistance to interference factors during the segmentation process, effectively reducing the impact of these factors on segmentation results. In summary, compared to network models that only add CSCA attention mechanism or only add a discriminator, the GU-Net model shows higher accuracy and robustness in the segmentation task of ISIC skin lesion images.

4.2.3. Comparative experiments with related methods

A comparison was conducted between GU-Net and four methods: (1) Convolutional Block Attention Module [28] that adds both spatial and channel attention mechanisms, (2) Attention U-Net [18] that adds only spatial attention mechanism, and (3) CA-Net [5] that adds spatial, channel, and scale attention mechanisms. (4) MALUNet [39] with four added attention mechanisms: DGA, IEA, CAB, and SAB. We retrained these three networks on the ISIC 2016, ISIC 2017, and ISIC 2018 skin lesion image datasets and compared them with GU-Net.

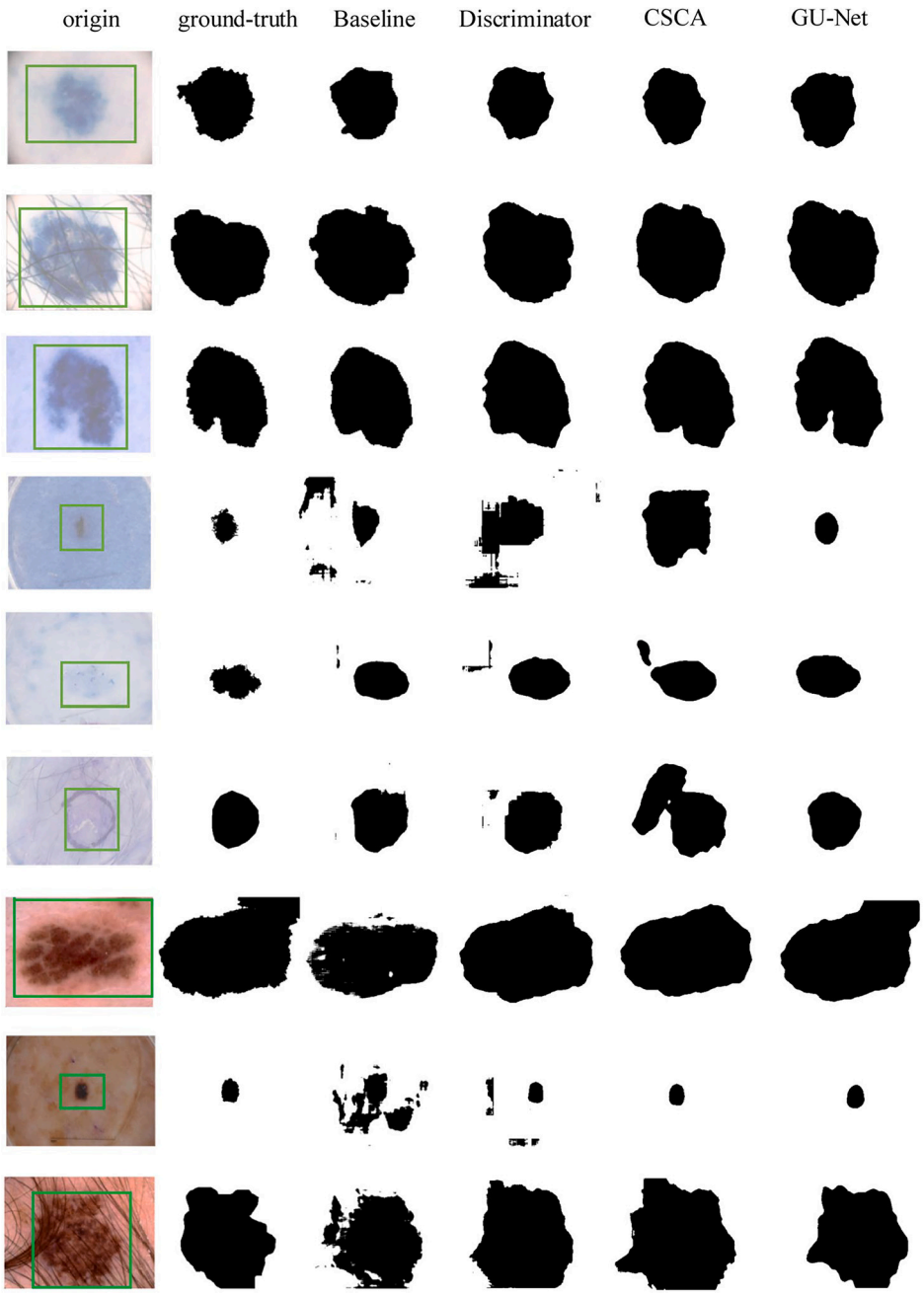

Fig. 5. Visual results of network models with different modules added, as well as the overall network, in the segmentation of ISIC skin lesion images.

The quantitative comparison of GU-Net with other methods for skin lesion image segmentation on the ISIC dataset is shown in Table 2. GU-Net outperforms other network models in terms of segmentation accuracy and performs equally well as other models in Dice score, ASSD, and IoU metrics. Specifically, when performing segmentation on the ISIC 2016, ISIC 2017, and ISIC 2018 datasets, GU-Net achieves Dice scores of 93.94%, 89.75%, and 92.42%, respectively, which is a significant improvement compared to U-Net with Dice scores of 92.75%, 86.82%, and 87.41%. Additionally, in comparison to the other three methods, GU-Net achieves the best accuracy in terms of the position information of the predicted results (IoU) for skin lesion image segmentation on the ISIC 2016, ISIC 2017, and ISIC 2018 datasets, with values of 88.98%, 82.41%, and 86.46%, respectively. Furthermore, GU-Net exhibits the smallest Average Symmetric Surface Distance (ASSD) with values of 0.4283pix, 1.1525pix, and 0.7057pix. In Table 2, we have added arrows indicating that higher/lower values are preferable, and bolded the best results. Through comparison, we find that combining the generative U-Net with the counterfactual attention mechanism based on channel and spatial dimensions, along with

Table 2
Quantitative comparison of GU-Net network with the baseline U-Net model and three other related network models incorporating different attention mechanisms in ISIC skin lesion image segmentation.

Datasets	Network	Dice(%)↑	ASSD(pix)↓	IoU(%)↑	Inference time(s)↓
ISIC2016	**Baseline**	92.75	0.5028	87.29	2.2051
	Attention U – Net	92.85	0.4465	87.30	2.1798
	CBAM	93.63	0.3683	88.61	**1.7216**
	CA – Net	93.46	0.6456	88.38	8.5925
	MALUNet	92.88	0.4389	87.34	3.7484
	GU – Net	**93.94**	**0.4283**	**88.98**	3.3289
ISIC2017	**Baseline**	86.82	1.6888	78.38	3.5597
	Attention U – Net	87.17	1.6922	78.80	4.0310
	CBAM	87.69	1.4839	79.61	**3.2568**
	CA – Net	88.53	1.5136	80.74	7.3255
	MALUNet	88.13	1.4317	78.78	6.1620
	GU – Net	**89.75**	**1.1525**	**82.41**	5.9833
ISIC2018	**Baseline**	87.41	1.0887	78.73	**3.0948**
	Attention U – Net	90.57	0.8716	83.47	5.1015
	CBAM	91.12	0.8179	84.37	4.3047
	CA – Net	91.94	0.7179	85.73	7.1549
	MALUNet	89.04	0.8871	80.20	4.5610
	GU – Net	**92.42**	**0.7057**	**86.46**	3.7532

the discriminator, improves the segmentation performance for skin lesion images. However, due to the need to train both the generator and the discriminator simultaneously, GU-Net has relatively longer inference time, particularly with large-scale datasets.

The visual results of GU-Net compared to other network models in ISIC skin lesion image segmentation are shown in Fig. 6. Specifically, we selected three representative groups of images from the experimental results on ISIC 2016, ISIC 2017, and ISIC 2018 datasets, which represent common challenges in skin lesion images, such as hair interference, discoloration, artificial markings, low contrast between lesion and skin, and blurred lesion edges. Each image consists of the original image from the publicly available ISIC skin lesion dataset (origin), the corresponding ground-truth, and the predicted segmentation results from different network architectures. The green box outlines the area of the lesion to be segmented, while the red box outlines the pigmented patches on the patient's skin itself. The first three images represent representative segmentation results under the conditions of blurred lesion edges, indistinct lesion boundaries due to low contrast with the patient's skin, and interference from color patches within the lesion region, respectively, in the ISIC 2016 dataset. The next three images represent representative segmentation results under the conditions of small lesion size with hair interference, blurred lesion edges with the presence of bubbles, hair, and markings, and low contrast between lesion and patient's skin, respectively, in the ISIC 2017 dataset. The last three images represent representative segmentation results under the conditions of blurred lesion edges with hair interference, areas with low contrast to the patient's skin within the lesion, and irregular lesion shapes with blurred edges, respectively, in the ISIC 2018 dataset.

After observing the comparison of visual results between GU-Net and U-Net, it is evident that incorporating attention mechanisms improves the accuracy of the network in ISIC skin lesion image segmentation tasks. Among them, the segmentation prediction results of GU-Net show significant advantages over other network models in several aspects. Firstly, GU-Net can more accurately identify the position of lesion areas, with its segmentation prediction results being closer to the actual location of the lesions. Secondly, when dealing with details, GU-Net demonstrates more meticulous and precise feature extraction capabilities, capturing more subtle changes and edge information, thereby achieving more accurate segmentation. Additionally, GU-Net with added attention mechanisms also exhibits strong resistance to interference factors, reducing the impact of interference factors on segmentation results, making the segmentation results more stable and reliable. In summary, GU-Net can generate segmentation results with the highest correlation to ground-truth in the task of ISIC skin lesion image segmentation, significantly improving segmentation accuracy.

4.3. Gland segmentation

4.3.1. Dataset

We used the Gland Segmentation (GlaS) dataset [40], which consists of microscopic images of Hema-toxylin and Eosin (H&E) stained slides, along with corresponding ground-truth annotations provided by pathologists. The dataset contains 165 images, which we randomly split into 85 for training and 80 for testing. Due to variations in image sizes within the dataset, we resized each image to a resolution of 128×128 pixels.

4.3.2. Experimental setup for ablation study

In this section, we perform quantitative and visual comparisons between GU-Net and network models that only include the CSCA attention mechanism, as well as those that only include a discriminator, on the GlaS gland dataset to investigate the effectiveness of the two modules. Here, CSCA refers to a network model that utilizes the channel and spatial-based counterfactual attention mechanism solely in the generator network, while Discriminator indicates a network model that incorporates the discriminator for interactive training.

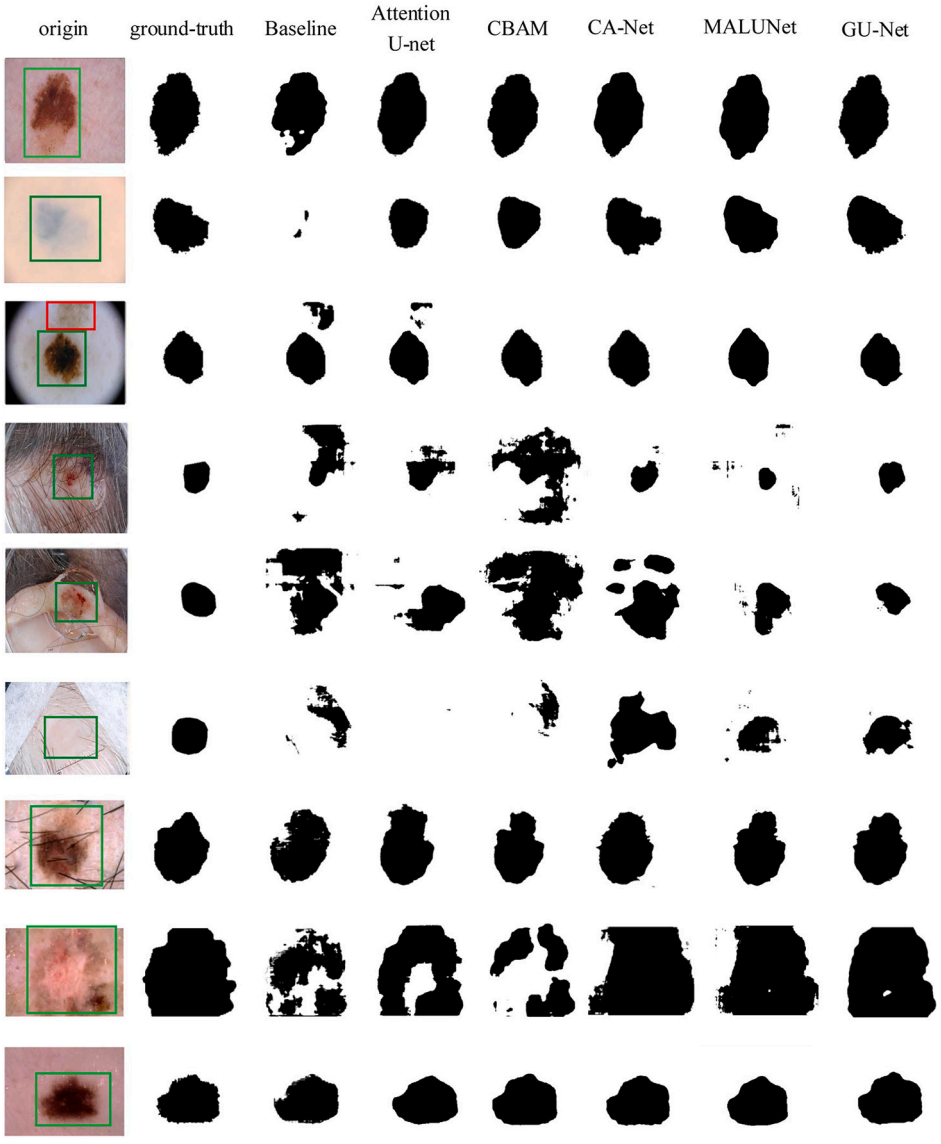

Fig. 6. Visual results of GU-Net and other network models in ISIC skin lesion image segmentation.

The quantitative comparison between the GU-Net network and network models that incorporate either only the CSCA attention mechanism or only the discriminator is shown in Table 3. The models are evaluated on the task of GlaS gland image segmentation. As evidenced in Table 3, the incorporation of both the CSCA and the discriminator resulted in enhanced performance when compared to the U-Net baseline model. Specifically, GU-Net significantly outperformed other variants in terms of Dice score, ASSD, and IoU, with corresponding values of 88.54%, 0.2053pix, and 79.62%. This demonstrates that GU-Net can enhance the segmentation performance of medical images. In Table 3, we have added arrows indicating that higher/lower values are preferable, and bolded the best results. However, due to the simultaneous training of the generator and discriminator in GU-Net, the inference time is longer compared to models with only the CSCA attention mechanism or the discriminator.

Fig. 7 illustrates the visual results of the U-Net base network model, a network model with only CSCA attention mechanism added, a network model with only the discriminator added, and the GU-Net network model in the segmentation of GlaS gland images. Each row of images consists of the original image (origin), corresponding ground-truth annotations, and predicted segmentation results of different network models from the GlaS gland dataset. From the visual results in the first row prediction images, it can be observed that when glandular lumina and stroma closely resemble glandular cell composition in gland tissue section images, the GU-Net network model can relatively accurately identify the position information of glandular cells and perform segmentation operations. Compared to the base network and models with only one module added, the segmentation results are closest to the corresponding ground-truth annotations. From the visual results in the second row prediction images, when glandular cells are numerous and dense, with significant morphological differences, the GU-Net network model provides the most detailed segmentation of small cells. Although it

Table 3
Quantitative comparison of network models with different added modules and the overall network in GlaS gland image segmentation task.

Network	Dice(%)↑	ASSD(pix)↓	IoU(%)↑	Inference time(s)↓
Baseline	83.04	0.2904	71.20	1.6149
Discriminator	87.33	0.2207	77.70	1.1778
CSCA	87.85	0.2144	78.58	**1.1701**
GU-Net	**88.54**	**0.2053**	**79.62**	1.1933

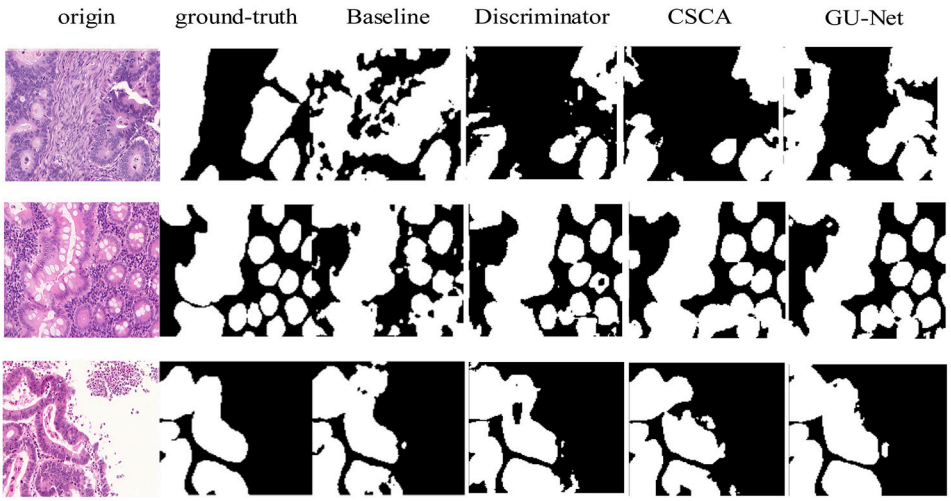

Fig. 7. Visualization results of adding different modules to the network models and the overall network for GlaS gland image segmentation.

may not identify smaller intercellular components, it can accurately delineate the contours of glandular cells with minimal influence from other cell types. From the visual results in the third row prediction images, when there is low contrast between intercellular stroma and glandular cell cytoplasm in gland tissue section images, and other interfering factors around the glands, compared to other network models, the GU-Net network model can more accurately differentiate glandular cell cytoplasm from interstitial components supporting and maintaining glandular structure, demonstrating more accurate segmentation performance.

In summary, the GU-Net network model exhibits higher accuracy in predicting the segmentation of glandular cells in cases where the original gland tissue images have extremely low contrast between the glandular cells and other structures, where there are numerous and densely packed glandular cells, where the shapes of glandular cells are unstable, and where the glandular cells contain regions similar to other components.

4.3.3. Comparative experiments with related methods

A comparison was conducted between GU-Net and four methods: (1) Convolutional Block Attention Module [28] that adds both spatial and channel attention mechanisms, (2)Attention U-Net [18] that adds only spatial attention mechanism, and (3) CA-Net [5] that adds spatial, channel, and scale attention mechanisms. (4) MALUNet [39] with four added attention mechanisms: DGA, IEA, CAB, and SAB. We retrained these three networks on the GlaS gland dataset and compared them with GU-Net.

The quantitative comparison of GU-Net with other methods for gland image segmentation on the GlaS dataset is shown in Table 4. GU-Net achieves a significant improvement with a Dice score of 88.54% while having the shortest inference time compared to U-Net, which has a Dice score of 83.04%. Furthermore, when compared to the other three methods, GU-Net achieves the lowest average symmetric surface distance (ASSD) and the highest accuracy in predicting the location information (IoU), with values of 0.2053pix and 79.62%, respectively, while consuming the shortest inference time. In Table 4, we have added arrows indicating that higher/lower values are preferable, and bolded the best results. Through comparison, it can be observed that combining the generative U-Net with channel and spatial dimension-based counterfactual attention mechanism along with a discriminator enhances the segmentation performance of the network for gland images.

The visual results of GU-Net compared to other network models in the glandular tissue image segmentation task of the GlaS dataset, are shown in Fig. 8. Each row of images, from left to right, includes the original tissue slice image from the GlaS gland dataset (origin), the corresponding ground-truth annotation, and the predicted segmentation result from different network models. By comparing the visualization results of the network models with added attention mechanisms to those of the U-Net baseline model, we observe that the addition of attention mechanisms enhances the accuracy of the network models in the GlaS glandular tissue image segmentation task. Specifically, the GU-Net network model demonstrates better segmentation performance when glandular cell regions contain areas similar to other structures or components, when the boundaries between adjacent glandular cells are small, and when the cells to be segmented have low contrast with the surrounding stroma.

Table 4
Quantitative comparison of GU-Net network with the baseline U-Net model and three other related network models with added attention mechanisms in GlaS gland image segmentation.

Network	Dice(%)↑	ASSD(pix)↓	IoU(%)↑	Inference time(s)↓
Baseline	83.04	0.2904	71.20	1.6149
Attention U-Net	82.62	0.2973	70.61	2.2701
CBAM	84.44	0.2701	73.25	1.2188
CA-Net	87.23	0.2362	77.47	2.4632
MALUNet	83.73	0.2879	72.76	4.2970
GU-Net	**88.54**	**0.2053**	**79.62**	**1.1933**

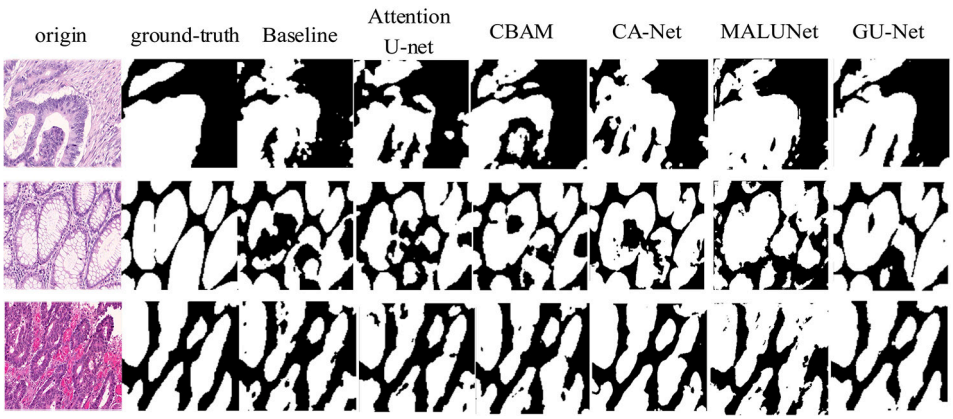

Fig. 8. Visualization results of GU-Net and other network models in GlaS gland image segmentation.

In the first row of the visualization results, it can be observed that the GU-Net network model can accurately identify the contours of glandular cells when there is low contrast and irregular shapes compared to the stroma, secretions, or other cells within the tissue slice, resulting in relatively accurate segmentation results compared to other network models. In the second row of the visualization results, the GU-Net network model demonstrates relatively accurate localization and morphology information when there are numerous glandular cells with stable shapes but indistinct boundaries between cells. In the third row of the visualization results, the GU-Net network model accurately identifies glandular cells of interest and extracts useful feature information for output when there is extremely low contrast between components such as glandular cytoplasm and structures like connective tissue, blood vessels, or other cells within the tissue slice, resulting in the highest overlap rate between the predicted segmentation map and the ground-truth annotated pixel points compared to other network models.

In summary, it can be observed from the visualization results that the predicted segmentation maps obtained by the GU-Net network model for the GlaS gland segmentation task are most similar to the corresponding ground-truth images. Thus, it is evident that the GU-Net network model exhibits the best segmentation performance for glandular cell segmentation tasks.

5. Discussion

For medical image segmentation tasks, the position, shape, and proportion of certain targets (such as lesion regions) can vary greatly. Additionally, due to the specific nature of medical images, there may be interference from factors related to the patient and the limited availability of data. Therefore, networks used for medical image segmentation need to understand and learn the positional relationships of the segmentation objects while reducing the impact of other interfering factors on the segmentation results. Convolutional neural networks generate feature maps with a large number of channels and often employ concatenation of feature maps with different semantic information. Focusing on the most relevant channels and spatial information is an effective approach to improving segmentation performance.

The CBAM attention mechanism is able to extract spatial information while focusing on the channel information of the feature maps. Therefore, we integrate the CBAM attention mechanism with the counterfactual attention mechanism to quantify the quality of attention by comparing the impact of facts (learned attention) and counterfactuals (unadjusted attention) on the final predictions (segmentation results). Then, by maximizing the difference, we encourage the network to learn more effective visual attention and reduce the influence of interfering factors. Additionally, GU-Net combines the improved generator network with the discriminator network in the GAN framework, where the updates to the generator network are not based on data samples but on the backpropagation from the discriminator. The generator and discriminator engage in an adversarial game through interactive training, enabling the generation of segmentation results with higher accuracy.

Like other deep learning models, the GU-Net model also requires a large amount of annotated data for training. However, in the field of medical image segmentation, obtaining a large quantity of high-quality publicly annotated datasets, especially for rare or

specific cases, is very challenging. To address this issue, we employed methods such as random image rotation during the experimental process to augment the data, which may lead to the problem of class imbalance in the input data. This imbalance may cause the GU-Net model to perform well on dominant classes but poorly on rare classes. Additionally, pathological tissues or cells exhibit a wide range of appearance representations, making it difficult for the GU-Net model to generalize to new types of lesions or rare pathological conditions that may not be adequately represented in the training data. Furthermore, deep neural networks are often considered black-box models, and the GU-Net model is no exception. Therefore, it is challenging to fully understand how and why it makes specific segmentation decisions for pathological tissues, organs, or cells.

Despite the limitations mentioned above, the GU-Net model has shown improved performance in medical image segmentation. We validated our approach on two representative image domains: the ISIC 2016, ISIC 2017, and ISIC 2018 skin lesion datasets, as well as the GlaS gland dataset. Experimental results demonstrate significant segmentation improvements of GU-Net compared to U-Net in both domains. In terms of Dice score, GU-Net achieved improvements of 1.19%, 2.93%, 5.01%, and 5.50% respectively. It also reduced ASSD by 0.0745pix, 0.5363pix, 0.3830pix, and 0.0851pix, and improved IoU by 1.69%, 4.03%, 7.73%, and 8.42%. These findings indicate that GU-Net exhibits competitive performance for different segmentation tasks in various modalities.

6. Conclusion

We propose the GU-Net network, which leverages causal reasoning to construct a generative U-Net network for medical image segmentation, aiming to improve the accuracy of the generated segmentation results. The generation network in GU-Net is built upon the U-Net base model and incorporates a counterfactual attention mechanism fused with the CBAM attention mechanism. By leveraging the causal relationship between predictions and attention, it intervenes on the learned visual attention to influence the network's predictions and maximize the segmentation performance, thereby encouraging the network to learn more useful feature information. The discriminator of the GAN network is employed as the discriminator in GU-Net, and it is jointly trained with the improved generator network through interactive training, enabling the generative U-Net network to generate more accurate segmentation feature maps.

The experimental results on skin original images demonstrate that GU-Net outperforms existing U-Net generative models with added attention mechanisms in terms of skin lesion image segmentation. Additionally, experiments on the GlaS gland dataset show that GU-Net achieves superior segmentation performance even in scenarios with small datasets, low contrast, and complex data with challenging segmentation tasks. This further validates the generalizability of GU-Net beyond a specific medical imaging domain, indicating its potential for future research and application in other fields.

CRediT authorship contribution statement

Dapeng Cheng: Writing – review & editing, Writing – original draft. **Jiale Gai:** Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology. **Bo Yang:** Writing – review & editing, Writing – original draft. **Yanyan Mao:** Writing – review & editing, Writing – original draft. **Xiaolian Gao:** Writing – review & editing. **Baosheng Zhang:** Writing – review & editing. **Wanting Jing:** Writing – review & editing, Writing – original draft. **Jia Deng:** Writing – review & editing, Writing – original draft. **Feng Zhao:** Writing – review & editing, Writing – original draft. **Ning Mao:** Writing – review & editing, Writing – original draft.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The datasets used in our experiments are publicly available for download, so we did not create a database for storage. Readers can search and download the relevant datasets from the official websites according to their needs and the respective regulations.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (62176140).

References

[1] M.E. Celebi, N. Codella, A. Halpern, Dermoscopy image analysis: overview and future directions, IEEE J. Biomed. Health Inform. 23 (2019) 474–478, https://doi.org/10.1109/JBHI.2019.2895803.

[2] J.C. Caicedo, A. Goodman, K.W. Karhohs, B.A. Cimini, J. Ackerman, M. Haghighi, C. Heng, T. Becker, M. Doan, C. McQuin, et al., Nucleus segmentation across imaging experiments: the 2018 data science bowl, Nat. Methods 16 (2019) 1247–1253.

[3] S. Ali, M. Dmitrieva, N.M. Ghatwary, S. Bano, G. Polat, A. Temizel, A. Krenzer, A. Hekalo, Y.B. Guo, B.J. Matuszewski, M. Gridach, I. Voiculescu, V. Yoganand, A. Chavan, A. Raj, N.T. Nguyen, D.Q. Tran, L.D. Huynh, N. Boutry, S. Rezvy, H. Chen, Y.H. Choi, A. Subramanian, V. Balasubramanian, X.W. Gao, H. Hu, Y. Liao, D. Stoyanov, C. Daul, S. Realdon, R. Cannizzaro, D. Lamarque, T. Tran-Nguyen, A. Bailey, B. Braden, J.E. East, J. Rittscher, Deep learning for detection and segmentation of artefact and disease instances in gastrointestinal endoscopy, Med. Image Anal. 70 (2021) 102002, https://doi.org/10.1016/j.media.2021.102002.

[4] D. Jha, S. Ali, H.D. Johansen, D.D. Johansen, J. Rittscher, M.A. Riegler, P. Halvorsen, Real-time polyp detection, localisation and segmentation in colonoscopy using deep learning, CoRR, arXiv:2011.07631 [abs], 2020, https://arxiv.org/abs/2011.07631, arXiv:2011.07631.
[5] R. Gu, G. Wang, T. Song, R. Huang, M. Aertsen, J. Deprest, S. Ourselin, T. Vercauteren, S. Zhang, Ca-net: comprehensive attention convolutional neural networks for explainable medical image segmentation, IEEE Trans. Med. Imaging 40 (2021) 699–711, https://doi.org/10.1109/TMI.2020.3035253.
[6] J.M.J. Valanarasu, P. Oza, I. Hacihaliloglu, V.M. Patel, Medical transformer: gated axial-attention for medical image segmentation, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Proceedings, Part I 24, Strasbourg, France, September 27–October 1, 2021, Springer, 2021, pp. 36–46.
[7] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015, IEEE Computer Society, 2015, pp. 3431–3440.
[8] O. Ronneberger, P. Fischer, T. Brox, U-net: convolutional networks for biomedical image segmentation, in: N. Navab, J. Hornegger, W.M. Wells III, A.F. Frangi (Eds.), Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015 - 18th International Conference Munich, Proceedings, Part III, Germany, October 5 - 9, 2015, in: Lecture Notes in Computer Science, vol. 9351, Springer, 2015, pp. 234–241.
[9] F. Milletari, N. Navab, S. Ahmadi, V-net: fully convolutional neural networks for volumetric medical image segmentation, in: Fourth International Conference on 3D Vision, 3DV 2016, Stanford, CA, USA, October 25-28, 2016, IEEE Computer Society, 2016, pp. 565–571.
[10] Ö. Çiçek, A. Abdulkadir, S.S. Lienkamp, T. Brox, O. Ronneberger, 3d u-net: learning dense volumetric segmentation from sparse annotation, in: S. Ourselin, L. Joskowicz, M.R. Sabuncu, G.B. Ünal, W.M. Wells III (Eds.), Medical Image Computing and Computer-Assisted Intervention - MICCAI 2016 - 19th International Conference, Proceedings, Part II, Athens, Greece, October 17-21, 2016, in: Lecture Notes in Computer Science, vol. 9901, 2016, pp. 424–432.
[11] X. Xiao, S. Lian, Z. Luo, S. Li, Weighted res-unet for high-quality retina vessel segmentation, in: 2018 9th International Conference on Information Technology in Medicine and Education (ITME), IEEE, 2018, pp. 327–331.
[12] X. Li, H. Chen, X. Qi, Q. Dou, C. Fu, P. Heng, H-denseunet: hybrid densely connected unet for liver and tumor segmentation from CT volumes, IEEE Trans. Med. Imaging 37 (2018) 2663–2674, https://doi.org/10.1109/TMI.2018.2845918.
[13] S. Mehta, E. Mercan, J. Bartlett, D.L. Weaver, J.G. Elmore, L.G. Shapiro, Y-net: joint segmentation and classification for diagnosis of breast biopsy images, in: A.F. Frangi, J.A. Schnabel, C. Davatzikos, C. Alberola-López, G. Fichtinger (Eds.), Medical Image Computing and Computer Assisted Intervention - MICCAI 2018 - 21st International Conference, Proceedings, Part II, Granada, Spain, September 16–20, 2018, in: Lecture Notes in Computer Science, vol. 11071, Springer, 2018, pp. 893–901.
[14] J.M.J. Valanarasu, V.A. Sindagi, I. Hacihaliloglu, V.M. Patel, Kiu-net: overcomplete convolutional architectures for biomedical image and volumetric segmentation, IEEE Trans. Med. Imaging 41 (2021) 965–976.
[15] J.M.J. Valanarasu, V.A. Sindagi, I. Hacihaliloglu, V.M. Patel, Kiu-net: towards accurate segmentation of biomedical images using over-complete representations, in: A.L. Martel, P. Abolmaesumi, D. Stoyanov, D. Mateus, M.A. Zuluaga, S.K. Zhou, D. Racoceanu, L. Joskowicz (Eds.), Medical Image Computing and Computer Assisted Intervention - MICCAI 2020 - 23rd International Conference, Proceedings, Part IV, Lima, Peru, October 4-8, 2020, in: Lecture Notes in Computer Science, vol. 12264, Springer, 2020, pp. 363–373.
[16] H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, X. Han, Y. Chen, J. Wu, Unet 3+: a full-scale connected unet for medical image segmentation, in: 2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020, IEEE, 2020, pp. 1055–1059.
[17] V. Iglovikov, A. Shvets, Ternausnet: U-net with VGG11 encoder pre-trained on imagenet for image segmentation, CoRR, arXiv:1801.05746 [abs], 2018, http://arxiv.org/abs/1801.05746, arXiv:1801.05746.
[18] O. Oktay, J. Schlemper, L.L. Folgoc, M.C.H. Lee, M.P. Heinrich, K. Misawa, K. Mori, S.G. McDonagh, N.Y. Hammerla, B. Kainz, B. Glocker, D. Rueckert, Attention u-net: learning where to look for the pancreas, CoRR, arXiv:1804.03999 [abs], 2018, http://arxiv.org/abs/1804.03999, arXiv:1804.03999.
[19] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A.L. Yuille, Semantic image segmentation with deep convolutional nets and fully connected crfs, arXiv preprint, arXiv:1412.7062, 2014.
[20] Z. Zhou, M.M.R. Siddiquee, N. Tajbakhsh, J. Liang, Unet++: a nested u-net architecture for medical image segmentation, in: D. Stoyanov, Z. Taylor, G. Carneiro, T.F. Syeda-Mahmood, A.L. Martel, L. Maier-Hein, J.M.R.S. Tavares, A.P. Bradley, J.P. Papa, V. Belagiannis, J.C. Nascimento, Z. Lu, S. Conjeti, M. Moradi, H. Greenspan, A. Madabhushi (Eds.), Deep Learning in Medical Image Analysis - and - Multimodal Learning for Clinical Decision Support - 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Proceedings, Granada, Spain, September 20, 2018, in: Lecture Notes in Computer Science, vol. 11045, Springer, 2018, pp. 3–11.
[21] A.M. Khan, A. Ashrafee, F.S. Khan, M.B. Hasan, M.H. Kabir, Attresdu-net: medical image segmentation using attention-based residual double u-net, in: International Joint Conference on Neural Networks, IJCNN 2023, Gold Coast, Australia, June 18-23, 2023, IEEE, 2023, pp. 1–8.
[22] O. Petit, N. Thome, C. Rambour, L. Themyr, T. Collins, L. Soler, U-net transformer: self and cross attention for medical image segmentation, in: C. Lian, X. Cao, I. Rekik, X. Xu, P. Yan (Eds.), Machine Learning in Medical Imaging - 12th International Workshop, MLMI 2021, Held in Conjunction with MICCAI 2021, Proceedings, Strasbourg, France, September 27, 2021, in: Lecture Notes in Computer Science, vol. 12966, Springer, 2021, pp. 267–276.
[23] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, L. Zhang, Bottom-up and top-down attention for image captioning and visual question answering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6077–6086.
[24] G. Chen, J. Lu, M. Yang, J. Zhou, Spatial-temporal attention-aware learning for video-based person re-identification, IEEE Trans. Image Process. 28 (2019) 4192–4205.
[25] Y. Rao, J. Lu, J. Zhou, Learning discriminative aggregation network for video-based face recognition and person re-identification, Int. J. Comput. Vis. 127 (2019) 701–718, https://doi.org/10.1007/s11263-018-1135-x.
[26] C. Guo, M. Szemenyei, Y. Yi, W. Wang, B. Chen, C. Fan, Sa-unet: spatial attention u-net for retinal vessel segmentation, in: 25th International Conference on Pattern Recognition, ICPR 2020, Virtual Event / Milan, Italy, January 10-15, 2021, IEEE, 2020, pp. 1236–1242.
[27] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18–22, 2018, Computer Vision Foundation / IEEE Computer Society, 2018, pp. 7132–7141, http://openaccess.thecvf.com/content_cvpr_2018/html/Hu_Squeeze-and-Excitation_Networks_CVPR_2018_paper.html.
[28] S. Woo, J. Park, J. Lee, I.S. Kweon, CBAM: convolutional block attention module, in: V. Ferrari, M. Hebert, C. Sminchisescu, Y. Weiss (Eds.), Computer Vision - ECCV 2018 - 15th European Conference, Proceedings, Part VII, Munich, Germany, September 8-14, 2018, in: Lecture Notes in Computer Science, vol. 11211, Springer, 2018, pp. 3–19.
[29] Z. Wang, N. Zou, D. Shen, S. Ji, Non-local u-nets for biomedical image segmentation, in: The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, the Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, the Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020, AAAI Press, 2020, pp. 6315–6322, https://ojs.aaai.org/index.php/AAAI/article/view/6100.
[30] Y. Rao, G. Chen, J. Lu, J. Zhou, Counterfactual attention learning for fine-grained visual categorization and re-identification, in: 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021, IEEE, 2021, pp. 1005–1014.
[31] J. Pearl, Direct and indirect effects, CoRR, arXiv:1301.2300 [abs], 2013, http://arxiv.org/abs/1301.2300, arXiv:1301.2300, 2013.
[32] K. Tang, Y. Niu, J. Huang, J. Shi, H. Zhang, Unbiased scene graph generation from biased training, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020, Computer Vision Foundation / IEEE, 2020, pp. 3713–3722, https://openaccess.thecvf.com/content_CVPR_2020/html/Tang_Unbiased_Scene_Graph_Generation_From_Biased_Training_CVPR_2020_paper.html.
[33] T. VanderWeele, Explanation in Causal Inference: Methods for Mediation and Interaction, Oxford University Press, 2015.

[34] I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A.C. Courville, Y. Bengio, Generative adversarial nets, in: Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, K.Q. Weinberger (Eds.), Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada, 2014, pp. 2672–2680, https://proceedings.neurips.cc/paper/2014/hash/5ca3e9b122f61f8f06494c97b1afccf3-Abstract.html.

[35] D.A. Gutman, N.C.F. Codella, M.E. Celebi, B. Helba, M.A. Marchetti, N.K. Mishra, A. Halpern, Skin lesion analysis toward melanoma detection: a challenge at the international symposium on biomedical imaging (ISBI) 2016, hosted by the international skin imaging collaboration (ISIC), CoRR, arXiv:1605.01397 [abs], 2016, http://arxiv.org/abs/1605.01397, arXiv:1605.01397.

[36] N.C.F. Codella, D.A. Gutman, M.E. Celebi, B. Helba, M.A. Marchetti, S.W. Dusza, A. Kalloo, K. Liopyris, N.K. Mishra, H. Kittler, A. Halpern, Skin lesion analysis toward melanoma detection: a challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (ISIC), in: 15th IEEE International Symposium on Biomedical Imaging, ISBI 2018, Washington, DC, USA, April 4-7, 2018, IEEE, 2018, pp. 168–172.

[37] N.C.F. Codella, V. Rotemberg, P. Tschandl, M.E. Celebi, S.W. Dusza, D.A. Gutman, B. Helba, A. Kalloo, K. Liopyris, M.A. Marchetti, H. Kittler, A. Halpern, Skin lesion analysis toward melanoma detection 2018: a challenge hosted by the international skin imaging collaboration (ISIC), CoRR, arXiv:1902.03368 [abs], 2019, http://arxiv.org/abs/1902.03368, arXiv:1902.03368.

[38] P. Tschandl, C. Rosendahl, H. Kittler, The HAM10000 dataset: a large collection of multi-source dermatoscopic images of common pigmented skin lesions, CoRR, arXiv:1803.10417 [abs], 2018, http://arxiv.org/abs/1803.10417, arXiv:1803.10417, 2018.

[39] J. Ruan, S. Xiang, M. Xie, T. Liu, Y. Fu, Malunet: a multi-attention and light-weight unet for skin lesion segmentation, in: D.A. Adjeroh, Q. Long, X.M. Shi, F. Guo, X. Hu, S. Aluru, G. Narasimhan, J. Wang, M. Kang, A. Mondal, J. Liu (Eds.), IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2022, Las Vegas, NV, USA, December 6-8, 2022, IEEE, 2022, pp. 1150–1156.

[40] K. Sirinukunwattana, J.P. Pluim, H. Chen, X. Qi, P.-A. Heng, Y.B. Guo, L.Y. Wang, B.J. Matuszewski, E. Bruni, U. Sanchez, et al., Gland segmentation in colon histology images: the glas challenge contest, Med. Image Anal. 35 (2017) 489–502.