The Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)

# Non-Local Context Encoder: Robust Biomedical Image Segmentation against Adversarial Attacks

**Xiang He,[1] Sibei Yang,[2] Guanbin Li,[1]* Haofeng Li,[2] Huiyou Chang,[1] Yizhou Yu[3]**
[1]School of Data and Computer Science, Sun Yat-sen University, China
[2]The University of Hong Kong, Hong Kong [3]Deepwise AI Lab, China
hexiang7@mail2.sysu.edu.cn, sbyang9@hku.hk, liguanbin@mail.sysu.edu.cn
lhaof@foxmail.com, isschy@mail.sysu.edu.cn, yizhouy@acm.org

## Abstract

Recent progress in biomedical image segmentation based on deep convolutional neural networks (CNNs) has drawn much attention. However, its vulnerability towards adversarial samples cannot be overlooked. This paper is the first one that discovers that all the CNN-based state-of-the-art biomedical image segmentation models are sensitive to adversarial perturbations. This limits the deployment of these methods in safety-critical biomedical fields. In this paper, we discover that global spatial dependencies and global contextual information in a biomedical image can be exploited to defend against adversarial attacks. To this end, non-local context encoder (NLCE) is proposed to model short- and long-range spatial dependencies and encode global contexts for strengthening feature activations by channel-wise attention. The NLCE modules enhance the robustness and accuracy of the non-local context encoding network (NLCEN), which learns robust enhanced pyramid feature representations with NLCE modules, and then integrates the information across different levels. Experiments on both lung and skin lesion segmentation datasets have demonstrated that NLCEN outperforms any other state-of-the-art biomedical image segmentation methods against adversarial attacks. In addition, NLCE modules can be applied to improve the robustness of other CNN-based biomedical image segmentation methods.

## Introduction

Biomedical image analysis catches people's eyes due to its popular application in computer-aided diagnosis and medical plan recommendation. Biomedical image segmentation is fundamental in biomedical image analysis, which performs pixel-level annotation for regions of interest (e.g. organs, substructures, and lesions) on biomedical images (e.g. X-ray, Magnetic Resonance Imaging, Computerized Tomography). However, it is challenging to obtain accurate segmentation because of the large shape and size variations of regions of interest, and the diversity of images produced by different biomedical imaging equipments (Hwang and Park 2017; Sarker et al. 2018). State-of-the-art biomedical image segmentation methods are based on fully convolutional networks (FCN) (Long, Shelhamer, and Darrell 2015), which is a type of deep convolutional neural networks (CNNs) designed for semantic segmentation in computer vision. The accuracy of CNN-based biomedical image segmentation has been beyond that of traditional ones (Litjens et al. 2017; Sarker et al. 2018; Hwang and Park 2017; Ronneberger, Fischer, and Brox 2015; Novikov et al. 2018; Yuan 2017). In addition to the accuracy of biomedical image segmentation, its stability and robustness are also essential for the fault-free clinical practice.

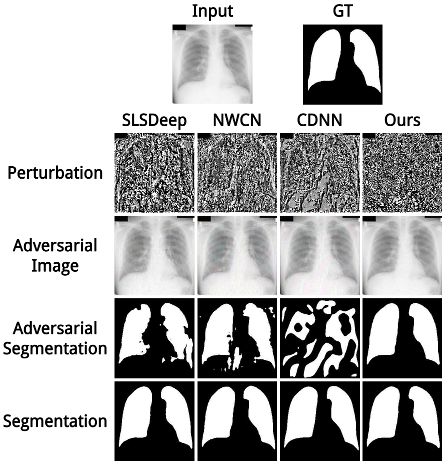

Figure 1: Sample adversarial attacks on SLSDeep (Sarker et al. 2018), NWCN (Hwang and Park 2017), CDNN (Yuan 2017) and our NLCEN. The input chest radiograph and its ground-truth segmentation are shown in the first row. Adversarial perturbations and images generated for models by the Iterative FGSM attack method (Kurakin, Goodfellow, and Bengio 2016) with adversarial intensity set to 16 are shown in the second and third rows respectively. Segmentation results on the adversarial images and the input image are shown in the fourth and fifth rows respectively.

Although CNN-based methods excel in solving many visual recognition tasks (LeCun, Bengio, and Hinton 2015; Ren et al. 2015; Li et al. 2017; Li and Yu 2018; Li et al.

*Xiang He and Sibei Yang contributed equally to this work. Corresponding author is Guanbin Li (Email: liguanbin@mail.sysu.edu.cn). This work was partially supported by the National Natural Science Foundation of China under Grant No.61702565 and the Fundamental Research Funds for the Central Universities under Grant No.18lgpy63.

Copyright © 2019, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

2018), the vulnerability of CNNs to adversarial attacks cannot be overlooked (Szegedy et al. 2013). Adversarial samples are legitimate samples with human-imperceptible perturbations, which attempt to fool a trained model to make incorrect predictions with high confidence (Szegedy et al. 2013). Such human-imperceptible perturbations, that CNNs are very sensitive to, are called adversarial noise. By exploiting the gradient-based error back-propagation mechanism for CNN training, adversarial attacks generate adversarial noise in an input image by back-propagating the error gradient induced by an intended incorrect prediction through a trained CNN model.

Recent work shows that complex semantic segmentation models, which are trained with an independent cross-entropy loss at each pixel on an image, are threatened by adversarial attacks (Xie et al. 2017; Arnab, Miksik, and Torr 2018). Although biomedical image segmentation models share a similar deep learning framework with semantic segmentation models, adversarial attacks targeted at them have not been well explored. Since biomedical image segmentation does not have sufficient high-quality training samples, trained models can easily experience overfitting and exhibit a weak generalization capability, which make them more sensitive to noise. This property makes the models more vulnerable when facing adversarial attacks, and challenges their use in safety-critical biomedical fields. An example of adversarial attacks on lung segmentation is shown in Figure 1.

The common defense strategy against adversarial attacks is adversarial training, which injects adversarial samples into training data to improve the robustness of trained models. Tramèr et al. 2018 show that if the adversarial samples are taken as augmented data, the consequence of adversarial attacks can be alleviated. However, this defense strategy is limited because the adversarial samples are obtained from specific models and their corresponding adversarial attack methods. Therefore, instead of a limited training strategy, we wish to design a generic module, which can be easily integrated into CNN-based biomedical image segmentation networks to improve their robustness.

The robustness of biomedical image segmentation can be improved effectively by global spatial dependencies and global contextual information. Therefore, we propose to model them with a module called non-local context encoder (NLCE). In order to better introduce the effectiveness of global spatial dependencies and global contexts, we use a single pixel as an example, and the situation of a single pixel can be easily extended to the entire image because segmentation models are trained with independent loss at every pixel. First, global spatial dependencies are very important in defending against adversarial attacks. Given a pixel, capturing its global spatial dependencies means finding all highly related pixels within the entire image, and the prediction at this pixel is affected by all those pixels. There are two perspectives to understand the effectiveness of global dependencies. One is that if an incorrect label was given to a pixel, the incorrect loss at the pixel would be passed to all other related pixels by back-propagation, which increases the intensity of perturbation, and makes the adversarial sample significantly different from the original image. The other is that the noise at a pixel can be gradually weakened by the fusion with its highly related pixels in the process of forward-propagation. Second, global contextual information has a positive effect in defending against adversarial attacks because the configuration of the human body is relatively stable. For example, in lung image segmentation, the left and right lungs provide geometric contextual information by learning their geometric relationship with respect to each other. Because of the association between the left and right lungs, the right lung needs to receive the same perturbation-based attacks when the left lung is attacked. Therefore, the intensity of the required perturbation is increased. Unfortunately, on one hand, CNNs have difficulty in capturing global dependencies because convolution operations only capture short-range dependencies by processing one local neighborhood at a time. Although stacked convolution operations are capable of capturing long-range dependencies by enlarging receptive fields (Fukushima 1980; Lecun et al. 1989), they increase the difficulty of optimization and may face the problem of gradient vanishing. On the other hand, current biomedical image segmentation methods do not make full use of global contextual information.

Inspired by the above analysis, in this paper, we propose a robust non-local context encoder module for biomedical image segmentation. The NLCE module captures the global spatial dependencies within a feature map by obtaining the response at a position of the feature map as a weighted sum of the features at all positions, and strengthens the features with channel-wise attention computed from the encoded global contextual information. In principle, the proposed robust NLCE module can also be applied to all CNN-based biomedical image segmentation methods and is able to improve the robustness of these models against adversarial attacks.

Moreover, we design and implement a medical image segmentation framework, named non-local context encoding network (NLCEN), which consists of two phases, the global phase and the refinement phase. Our global network is based on the feature pyramid network (FPN) (Lin et al. 2017) and our NLCE modules. It learns global feature representations at different levels. The refinement network fuses features at different levels to obtain sharp boundaries. We conduct experiments on two common benchmark biomedical image segmentation datasets, the JSRT dataset for lung segmentation (Shiraishi et al. 2000) and the ISBI 2016 dataset (Gutman et al. 2016) for skin lesion segmentation. Experimental results show that our NLCEN with NLCE modules has both high segmentation accuracy and robustness against adversarial attacks, and the NLCE modules practically help improve the segmentation accuracy of other biomedical image segmentation methods when they face adversarial attacks.

In summary, this paper has the following contributions:

- This is the first paper, to the best of our knowledge, attempts to improve the robustness of biomedical image segmentation methods by adding a robust module to the network. It proposes to exploit global spatial dependencies and global contexts to effectively improve the robustness of biomedical image segmentation methods.

- It proposes non-local context encoder (NLCE), which is a robust biomedical image segmentation module against adversarial attacks. The NLCE module is able to capture distance-independent dependencies and global contextual information. And it can be easily applied to other CNN-based image segmentation methods.
- It introduces non-local context encoding network (NLCEN), which achieves high segmentation accuracy and is robust on adversarial samples with different levels of adversarial perturbations.

## Related Work

### Biomedical Image Segmentation

The state-of-the-art biomedical image segmentation methods have similar frameworks to CNNs-based semantic segmentation models, but with fewer convolutional blocks and fewer network parameters to avoid overfitting. The U-net network is the most well-known segmentation method for biomedical image segmentation, and it is based on FCNs, but its upsampling phase and the downsampling phase use the same number of convolution operations in each level and the skip connection is used to connect the downsampling layer to the upsampling layer (Ronneberger, Fischer, and Brox 2015). InvertedNet is an improved version of U-net that has fewer parameters to reduce overfitting, and for more accurate localization, it adopts delayed subsampling and learns higher resolution features (Novikov et al. 2018). In order to use contextual information while maintaining resolution, NWCN adopts an atrous convolution-based model and utilizes a multi-stage training strategy to refine the preliminary segmentation results (Hwang and Park 2017). CDNN is also based on FCNs and it designs a loss function based on Jaccard distance (Yuan 2017). SLSDeep, consisting of skip-connections, dilated residual and pyramid pooling, is an efficient skin lesion segmentation model from dermoscopic images. Its loss function, including negative log likelihood and end point error loss, is designed to obtain sharp boundary (Sarker et al. 2018).

### Adversarial Attacks

Since the adversarial attacks to deep neural networks have been proposed by Szegedy et al., they have received extensive attention. They are designed to generate the adversarial samples to fool a trained model to make incorrect predictions with high confidence. The adversarial perturbation is estimated by solving penalized optimization problem by using L-BFGS optimization method (Szegedy et al. 2013). Goodfellow, Shlens, and Szegedy believe that the main reason why neural networks are vulnerable to adversarial attack is their linear behavior in high-dimensional space and propose the single-step fast gradient sign method (FGSM) to generate adversarial samples directly and efficiently. The single-step targeted attack is a modified version of FGSM, which aims at reducing the loss function of target category instead of the increasing the loss function of the original category (Kurakin, Goodfellow, and Bengio 2016). In addition, the proposed basic iterative method can increase the success rate of attacks. The adversarial samples generated by iterative methods are less transferable than those generated by single-step attacks (Kurakin, Goodfellow, and Bengio 2016; Arnab, Miksik, and Torr 2018). Xie et al. are the first to explore adversarial attacks on image segmentation and detection on large datasets and propose the density adversary generation to generate effective adversarial samples by considering all the targets simultaneously. Arnab, Miksik, and Torr present the first rigorous evaluation on the robustness of the state-of-the-art semantic segmentation models to single-step adversarial attacks and iterative adversarial attacks.

### Global Modeling

The global information modeling of images is an important part of the visual recognition field, and global information is utilized in many visual recognition tasks, e.g. scene segmentation (Li et al. 2016), saliency detection (Li and Yu 2016; Li et al. 2017) and semantic segmentation (Zhang et al. 2018). Getting global image information for CNN-based models is challenging, and it needs to consider both local dependencies and long-range dependencies. Stacked convolutional blocks can only capture local information due to restricted receptive fields. LSTM-CF treats spatial feature map obtained by CNNs as horizontal and vertical sequences respectively. It adopts multiple bi-directional long short term memory networks (LSTMs) in vertical direction to capture vertical short and long-range context, then the context is fused to get global spatial information by applying another bi-directional LSTMs in horizontal (Li et al. 2016). However, recurrent operations, like LSTMs, are still progress a local neighbor at a time, and the connection between two distant points must pass through the intermediate points. To capture the long-distance dependency, (Wang et al. 2018) proposed a fast and direct method, which considers the features at all the positions to capture the dependencies at a position in a low-level feature map. Zhang et al. takes the entire dataset into account and learns a set of global inherent representative of features to capture the global context for images. The global information for a feature map is obtained by encoding the relationships between its all features and the representative features.

## Methodology

We propose a robust biomedical image segmentation module, called non-local context encoder (NLCE), against adversarial attacks. NLCEs capture short- and long-range spatial dependencies and strengthen the features with channel-wise feature map attention using the encoded global contexts. The effectiveness of global spatial dependencies and global contextual information contributes to the robustness of NLCE against attacks. In order to refine segmentation and capture sharp boundaries, we introduce coarse-to-fine non-local context encoding network (NLCEN), which captures the robust enhanced feature representations at different levels and then learns the fused multi-scale features. In this section, we introduce the NLCE module and the NLCEN framework in more detail.

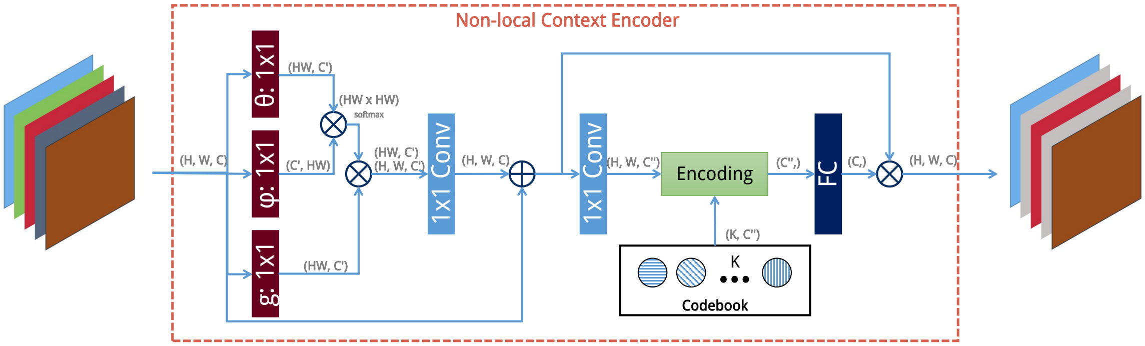

Figure 2: The architecture of our proposed non-local context encoder (NLCE). Our NLCE module first enhances and denoises the feature map by modeling global spatial dependencies and then applies channel-wise feature map attention by using encoded global context computed from a learned codebook.

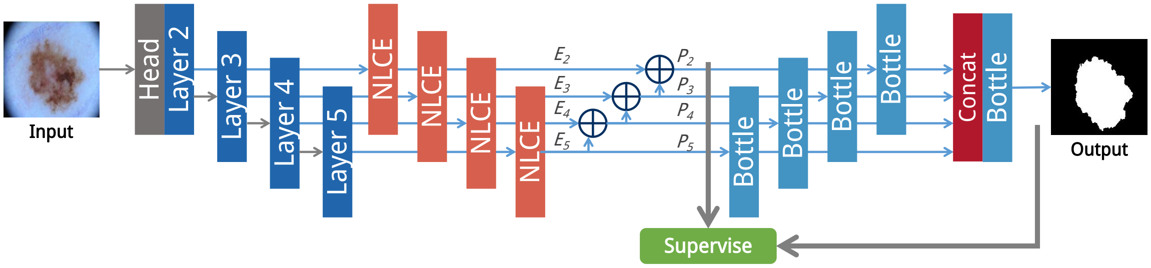

Figure 3: The overall architecture of our proposed non-local context encoding network (NLCEN). The left part is based on a ResNet backbone and a feature pyramid. An NLCE module is added to bottom-up feature activations before lateral connections at different levels, and independent supervision is applied to predictions at all levels. The multi-scale information fused from all the pyramid features are used to refine the prediction and produce segmentation.

## Non-Local Context Encoder

Our non-local context encoder takes an $H \times W \times C$ feature map as input. It captures spatial short- and long-range dependencies in the feature map by following the design by Wang et al. 2018. It considers the feature map as a set of $C$-dimensional features $X = \{\mathbf{x}_1, ...\mathbf{x}_N\}$, where $N = H \times W$ is the total number of features. We define the pairwise function $f$ that learns a relationship between any two features $\mathbf{x}_i$ and $\mathbf{x}_j$ as

$$f(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(\theta(\mathbf{x}_i)^T \phi(\mathbf{x}_j)\right), \tag{1}$$

where $\theta(\mathbf{x}_i) = W_\theta \mathbf{x}_i$ and $\phi(\mathbf{x}_j) = W_\phi \mathbf{x}_j$ are feature embeddings, where $W_\theta$ and $W_\phi$ are learned weight matrices.

The non-local response $\mathbf{y}_i$ for feature $\mathbf{x}_i$ is defined as

$$\mathbf{y}_i = \frac{1}{C(\mathbf{x})} \sum_{j=1}^{N} f(\mathbf{x}_i, \mathbf{x}_j) g(\mathbf{x}_j), \tag{2}$$

where the unary function $g$ is a mapping with a learned weight matrix $W_g$ to compute the representation $g(\mathbf{x}_j) = W_g \mathbf{x}_j$ of $\mathbf{x}_j$. $C(\mathbf{x})$ is the normalization factor, defined as $C(\mathbf{x}) = \sum_{i=1}^{N} f(\mathbf{x}_i, \mathbf{x}_j)$. The non-local response $\mathbf{y}_i$ captures short- and long-range dependencies via considering all features in the above non-local operation.

Next, the enhanced features $\mathbf{z}_i = W_z \mathbf{y}_i + \mathbf{x}_i$ ($W_z$ maps $\mathbf{y}_i$ to the $C$-dimensional space), which combine the non-local response $\mathbf{y}_i$ with the original feature $\mathbf{x}_i$, are fed into the context encoder discussed below. The feature map with the size of $H \times W \times C$ constructed from the enhanced features is denoted as $F_z$. Inspired by Zhang et al. 2018, we learn a global codebook $D = \{\mathbf{d}_1, ...\mathbf{d}_K\}$, which contains $K$ $C''$-dimensional codewords. The codebook represents global statistical information about the non-local enhanced features, and each codeword represents a visual center. We transform the enhanced features to the same dimensionality as the codewords via a $1 \times 1$ convolution, and the resulting $C''$-dimensional features are denoted as $Z' = \{\mathbf{z}'_1, ...\mathbf{z}'_N\}$. The normalized residual $\mathbf{e}_{ik}$ between an enhanced feature $\mathbf{z}'_i$ and a codeword $\mathbf{d}_k$ is defined as

$$\mathbf{e}_{ik} = \frac{\exp\left(-s_k \|\mathbf{r}_{ik}\|^2\right)}{R(\mathbf{e}_i)} \mathbf{r}_{ik}, \tag{3}$$

where $\mathbf{r}_{ik} = \mathbf{z}'_i - \mathbf{d}_k$ is the residual between feature $\mathbf{z}'_i$ and codeword $\mathbf{d}_k$, $s_k$ is a learned smoothing factor for codeword $\mathbf{d}_k$, and $R(\mathbf{e}_i) = \sum_{l=1}^{K} \exp(-s_l \|\mathbf{r}_{il}\|^2)$ is the normalization factor for feature $\mathbf{x}_i$. Thus, the residual information for all features captured by the codeword $\mathbf{d}_k$ is defined as $\mathbf{e}_k = \sum_{i=1}^{N} \mathbf{e}_{ik}$, and the global context is defined as $\mathbf{e} = \sum_{k=1}^{K} \sigma(\mathbf{e}_k)$, where $\sigma$ denotes Batch Normalization with ReLU.

Then, the global context $\mathbf{e}$ encoded from the spatial non-local features is used to strengthen the features using channel-wise feature map attention by predicting a channel-wise scaling factor $\gamma = \text{sigmoid}(W_\gamma \mathbf{e})$, where $W_\gamma$ is a learned weight matrix. The output from the NLCE module, $F_z \otimes \gamma$, is a channel-wise multiplication between the non-local enhanced feature map $F_z$ and the channel-wise scaling factor $\gamma$.

The architecture of our NLCE module is shown in Figure 2. The NLCE module first captures short- and long-range spatial dependencies to denoise and strengthen the feature map, and then scales the feature map channels by scaling factors predicted using the encoded global context. Global dependencies and global contexts reduce the negative impact of adversarial noise, and give rise to the robustness of the NLCE module against adversarial attacks. Fusing information from the highly related pixels or the global context in forward propagation gradually weakens adversarial noise to a pixel or a semantic proposal.

## Non-Local Context Encoding Network (NLCEN)

Our proposed coarse-to-fine non-local context encoding network (NLCEN) takes one biomedical image as input and produces a segmentation of organs or lesions at the pixel level. NLCEN has two phases, and its overall architecture is shown in Figure 3.

The architecture of the global phase is based on the ResNet backbone (He et al. 2016) and feature pyramid network. The fused information of low-level and high-level features by upsampling high-level features can capture rich contextual information with high resolution. An NLCE module is attached to the last residual block of conv2 through conv5 respectively to obtain multi-level robust non-local feature maps, denoted as $E_2, ..., E_5$. Following FPN, the fused feature map $P_i$ $(i = 2, 3, 4)$ is obtained by element-wise addition between $E_i$ and the $1 \times 1$ convolved and upsampled $P_{i+1}$, and $P_5$ is obtained by attaching a $1 \times 1$ convolutional layer to $E_5$. Feature maps $P_2, ..., P_5$ are used to independently produce segmentation results by feeding each of them through a distinct $3 \times 3$ convolution filter and a bilinear interpolation layer. Supervision is directly applied to each of these segmentation results.

Multi-level feature maps are fused together via upsampling and concatenation after going through bottleneck operations (He et al. 2016), and the refined segmentation prediction is produced directly from the fused feature map with multi-scale information. The number of bottleneck operations is respectively $0, 1, 2, 3$ for feature maps $P_2, P_3, P_4, P_5$. During testing, the final output is produced from the refined segmentation prediction.

The loss function for a single map prediction is defined as the sum of cross-entropy losses at individual pixels between the ground truth and the predicted segmentation map:

$$L_s = \sum_{i}^{|I|} \log p_{i,g_i}, \tag{4}$$

where $|I|$ denotes the total number of pixels, $g_i$ is the ground-truth label at pixel $i$, $p_{i,g_i}$ is the probability that pixel $i$ is classified to category $g_i$.

We denote the loss for the segmentation predictions obtained from $P_2, ..., P_5$ as $L_g^2, ..., L_g^5$, and the loss for the refined segmentation as $L_r$. The total loss is defined as:

$$L = \frac{1}{4} \sum_{i=2}^{5} L_g^i + \lambda L_r, \tag{5}$$

where $\lambda = 0.25$ is a weight balancing multiple coarse predictions from the global phase and the refined prediction from the refinement phase.

# Experimental Results

## Datasets

We have conducted evaluations on two commonly used benchmark biomedical image datasets, the Japanese Society of Radiological Technology (JSRT) dataset for lung segmentation (Shiraishi et al. 2000) and the International Symposium on Biomedical Imaging (ISBI 2016) dataset for skin lesion segmentation (Gutman et al. 2016).

The JSRT dataset was first introduced to help diagnostic training and testing for tuberculosis. It contains 154 nodule and 93 non-nodule post-anterior (PA) chest radiographs with a $2048 \times 2048$ high resolution and wide density range. We split chest radiographs into a training set of 124 images and a test set of 123 images by following previous practices in the literature (Hwang and Park 2017). The ground truth for the JSTR dataset is provided in (Van Ginneken, Stegmann, and Loog 2006).

The ISBI 2016 dataset provides 900 training images and 379 testing images with binary masks of skin lesion. The size of the images ranges from $524 \times 718$ to $2848 \times 4288$.

## Adversarial Attacks

We adopt the target Iterative FGSM attack method (Kurakin, Goodfellow, and Bengio 2016) to generate adversarial samples for a concrete model because the iterative white-box attacking methods have a high success rate. An attack sets the target as the inverse of ground-truth masks, denoted as $S_t$, and the adversarial sample of a single example in each iteration is defined as:

$$\mathbf{x}_{t+1}^{adv} = \text{clip}(\mathbf{x}_t^{adv} - \alpha \cdot \text{sign}(\nabla_{\mathbf{x}_t^{adv}} L_r(f(\mathbf{x}_t^{adv}; \theta_f)), \epsilon), \tag{6}$$

where $\mathbf{x}_0^{adv}$ is initialized to $\mathbf{x}$, the intensity of the adversarial perturbation is $\epsilon$, the step size of iterations is denoted as $\alpha$, and $\theta_f$ represents network parameters.

Following Kurakin, Goodfellow, and Bengio, we set $\alpha = 1$, the number of iterations to $\min(\epsilon + 4, \lceil 1.25\epsilon \rceil)$, and the $L_\infty$ norm of adversarial perturbation to intensity. We generate adversarial samples by setting adversarial intensity to every value from $\{0.5, 1, 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32\}$.

## Evaluation Metrics

We evaluate the robustness of biomedical image segmentation methods by measuring the drop in segmentation accuracy after adding adversarial perturbations with different intensities to the original testing images. Dice's coefficient

($DIC$) and Jaccard similarity coefficient ($JSC$) are commonly used accuracy metrics in biomedical image segmentation. $DIC$ and $JSC$ are computed as follows:

$$DIC = \frac{2 \cdot TP}{2 \cdot TP + FN + FP}, \tag{7}$$

$$JSC = \frac{TP}{TP + FN + FP}, \tag{8}$$

where $TP$, $TN$, $FP$, $FN$ are the number of pixel-level true positives, true negatives, false positives, and false negatives, respectively.

## Implementation

Our proposed NLCEN with NLCE modules has been implemented on the open source deep learning framework, PyTorch(Paszke et al. 2017). We follow the same experimental setups as in Hwang and Park and Sarker et al.. Horizontal flips, vertical flips and random rotations with ±10 degrees are used as data augmentation operations on the ISBI 2016 dataset while no data augmentation is applied to the JSRT dataset during training. We set the mini-batch size to 8, and all input images are resized to $256 \times 256$. The Adam optimizer is adopted to update network parameters with the learning rate set to $0.001$ initially and reduced by $10\%$ whenever the training loss stops decreasing until $0.0001$. We use a weight decay of $0.0001$ and an exponential decay rate for the first moment estimates and the second moment estimates of $0.9$ and $0.999$ respectively. It takes 2 hours to train a model on the JSRT dataset in a single NVIDIA TITAN GPU and 2 more hours to generate adversarial samples for testing when an intensity of adversarial perturbation is given. The training and testing times on the ISBI 2016 dataset are $4$ hours respectively.

## Comparison with the State of the Art

We compare the robustness of our proposed NLCEN with that of five state-of-the-art methods for lung segmentation and skin lesion segmentation, including dilated residual and pyramid pooling networks (SLSDeep) (Sarker et al. 2018), network-wise training of convolutional networks (NWCN) (Hwang and Park 2017), convolutional networks for biomedical image segmentation (UNet) (Ronneberger, Fischer, and Brox 2015), fully convolutional architectures for multi-class segmentation (InvertNet) (Novikov et al. 2018) and segmentation with fully convolutional-deconvolutional networks (CDNN) (Yuan 2017). All the evaluations of the above networks are conducted on both the JSRT and ISBI 2016 datasets. According to the scale of the datasets, we adopt a ResNet-18 backbone for the JSRT dataset and a ResNet-50 backbone for the ISBI 2016 dataset. On each dataset, we first train a benchmark segmentation model on the training set and compute segmentation accuracy metrics ($DIC$ and $JSC$) on the testing set; and then, under each given intensity of perturbation, we generate adversarial samples of the testing set on the basis of the benchmark model and test its segmentation accuracy on the generated adversarial samples.

**Quantitative Evaluation** Figures 4 and 5 show evaluation results in terms of $DIC$ and $JSC$ on the JSRT and ISBI 2016 datasets respectively. In these figures, we can find that NLCEN achieves the highest accuracy on clean skin lesion images, and achieves almost the top performance on clean lung images. Even when the strongest adversarial perturbation ($\epsilon = 32$) is exerted, it still maintains the highest accuracy. Its accuracy drops by only $0.01$ ($0.971$ to $0.963$) in $DIC$ and $0.02$ ($0.945$ to $0.929$) in $JSC$ on the JSRT dataset, and drops by $0.11$ ($0.907$ to $0.801$) in $DIC$ and $0.14$ ($0.844$ to $0.704$) in $JSC$ on the ISBI 2016 dataset. The results show that adversarial attacks have almost no effects on our lung segmentation model. The drop in accuracy on the ISBI 2016 skin dataset is larger than that on the JSRT dataset because there is very little contextual information in skin lesion images. Even though, our NLCEN is still the most robust one of all the models. Moreover, this experiment also indicates that the outstanding robustness of our model against adversarial samples with different levels of perturbation intensity.

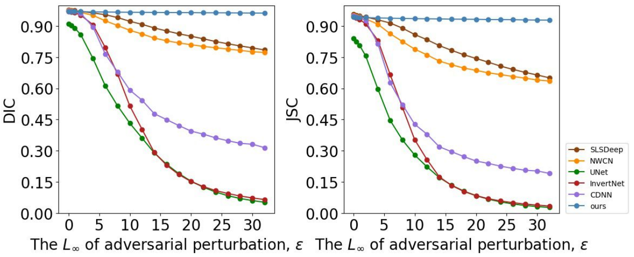

Figure 4: Comparison of quantitative results in terms of $DIC$ and $JSC$ on the JSRT lung segmentation dataset.

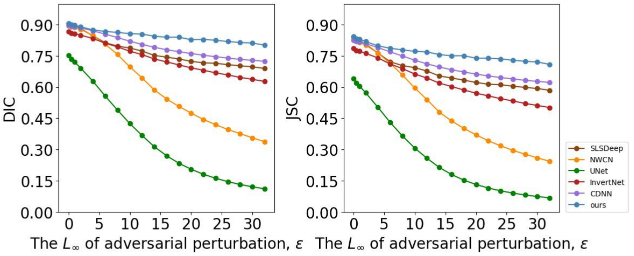

Figure 5: Comparison of quantitative results in terms of $DIC$ and $JSC$ on ISBI 2016 skin lesion segmentation dataset.

**Qualitative Evaluation** Figure 6 visually compares segmentation results from our model and five existing methods when they are under the attack of targeted Iterative FGSM with $\epsilon = 32$. Refer to the supplementary materials for more results.

Our method achieves the most accurate segmentations among all the methods when they face adversarial attacks. Segmentation results from our method on adversarial samples are almost the same as those of our method on clean images, which demonstrates the robustness of our model against adversarial attacks. Under adversarial attacks, our method produces accurate segmentations of lung images, and segments out skin lesions completely; on the other hand, lung segmentations obtained from NWCN, UNet, InvertNet

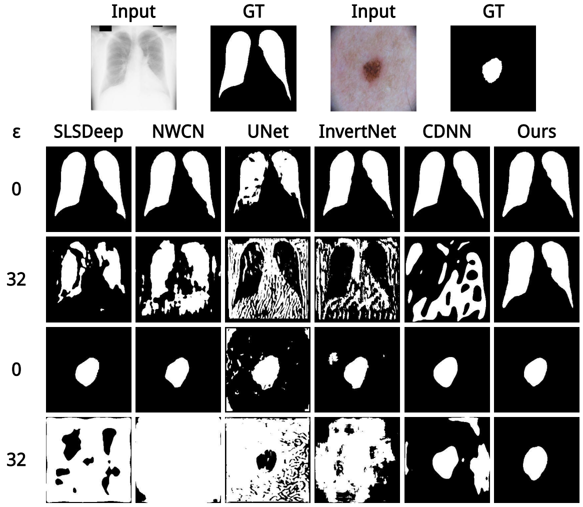

Figure 6: Comparison of segmentation results obtained from SLSDeep, NWCN, UNet, InvertNet, CDNN and our NLCEN when they are attacked by targeted Iterative FGSM with $\epsilon = 32$.

and CDNN are appalling, and all the other methods fail on skin lesion segmentation.

## Ablation Studies

As discussed in the Methodology section, the robustness of our NLCE modules against adversarial attacks comes from global spatial dependencies and global contextual information. To verify their validity and necessity, we compare NLCEN with its three variants (i.e. NLCEN without NLCE modules (w/o NLCE), NLCEN without modeling global dependencies (w/o NL) and NLCEN without capturing global contexts (w/o CE)), which are trained and tested on the JSTR dataset. For the fairness of the comparison, we train the w/o NLCE model first. Then, we fine-tune the w/o NL, w/o CE and NLCEN models separately by freezing the layers of the w/o NLCE model. Finally, we fine-tune NLCEN without freezing any layer to obtain the fine-tuned model.

The robustness of these models are evaluated and the results are shown in Figure 7. The non-local dependencies part or the global context part alone can already improve robustness, and the complete NLCE module with both parts can enhance the robustness further. That demonstrates the necessity of global dependencies and global contexts as well as the possibility of cooperation between them. In addition, the accuracy and robustness can be further enhanced by finetuning the NLCEN without freezing any layer.

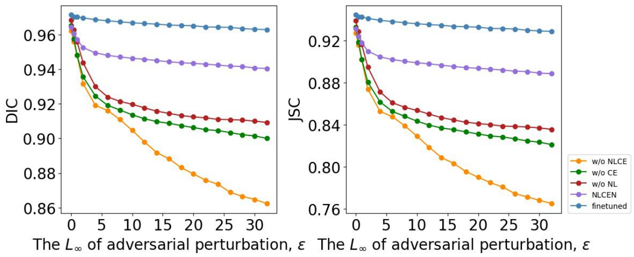

Figure 7: Ablation study on our non-local context encoding network.

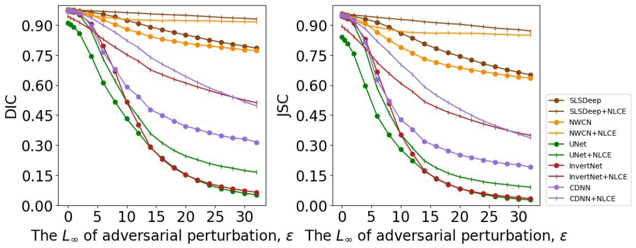

Figure 8: Comparison of robustness with and without nonlocal context encoder in other biomedical image segmentation methods.

## Generalization

To verify that our non-local context encoder can be easily integrated into other networks, we instantiate an NLCE version for SLSDeep, NWCN, UNet, InvertNet and CDNN networks, respectively. Except for NWCN, we add one NLCE module between the last downsampling layer and the first upsampling layer in each network. For NWCN, we add two NLCE modules because it has two subnetworks. Finally, we train these updated networks from scratch and test those networks with NLCE modules on the JSRT dataset.

Figure 8 shows a comparison between methods with NLCE modules and those without on the JSRT dataset. Methods with NLCE modules achieve significantly higher $DIC$ and $JSC$ than those without. This reveals NLCE modules are compatible with other biomedical image segmentation methods to strengthen their defense against adversarial attacks.

# Conclusions

In this paper, we have proposed a non-local context encoder which is a robust biomedical image segmentation module against adversarial attacks. It is designed to not only capture global spatial dependencies by learning the response at a single feature as a weighted sum of all the features, but also strengthen the features with channel-wise feature map attention by using encoded global contextual information. The NLCE modules are core components of our non-local context encoding network (NLCEN) for robust and accurate biomedical image segmentation. Experimental results on both lung segmentation and skin lesion segmentation datasets have demonstrated that our proposed method can denoise adversarial perturbations and defend against adversarial attacks effectively while achieving accurate segmentation. In addition, our NLCE modules can help improve the robustness of other biomedical image segmentation methods against adversarial attacks.

# References

Arnab, A.; Miksik, O.; and Torr, P. H. 2018. On the robustness of semantic segmentation models to adversarial attacks. In *Proceedings of CVPR*.

Fukushima, K. 1980. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics* 36(4):193–202.

Goodfellow, I.; Shlens, J.; and Szegedy, C. 2015. Explaining and harnessing adversarial examples. In *Proceedings of ICLR*.

Gutman, D.; Codella, N. C. F.; Celebi, M. E.; Helba, B.; Marchetti, M. A.; Mishra, N. K.; and Halpern, A. 2016. 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* abs/1605.01397.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of CVPR*, 770–778.

Hwang, S., and Park, S. 2017. Accurate lung segmentation via network-wise training of convolutional networks. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer. 92–99.

Kurakin, A.; Goodfellow, I.; and Bengio, S. 2016. Adversarial machine learning at scale. In *Proceedings of ICLR*.

LeCun, Y.; Bengio, Y.; and Hinton, G. 2015. Deep learning. *nature* 521(7553):436.

Lecun, Y.; Boser, B. E.; Denker, J. S.; Henderson, D.; Howard, R. E.; Hubbard, W.; and Jackel, L. D. 1989. Backpropagation applied to handwritten zip code recognition. *Neural Computation* 1(4):541–551.

Li, G., and Yu, Y. 2016. Visual saliency detection based on multiscale deep cnn features. *IEEE Transactions on Image Processing* 25(11):5012–5024.

Li, G., and Yu, Y. 2018. Contrast-oriented deep neural networks for salient object detection. *IEEE Transactions on Neural Networks and Learning Systems*.

Li, Z.; Gan, Y.; Liang, X.; Yu, Y.; Cheng, H.; and Lin, L. 2016. Lstm-cf: Unifying context modeling and fusion with lstms for rgb-d scene labeling. In *Proceedings of ECCV*, 541–557. Springer.

Li, G.; Xie, Y.; Lin, L.; and Yu, Y. 2017. Instance-level salient object segmentation. In *Proceedings of CVPR*, 247–256.

Li, G.; Gan, Y.; Wu, H.; Xiao, N.; and Lin, L. 2018. Cross-modal attentional context learning for rgb-d object detection. *IEEE Transactions on Image Processing*.

Lin, T.-Y.; Dollár, P.; Girshick, R. B.; He, K.; Hariharan, B.; and Belongie, S. J. 2017. Feature pyramid networks for object detection. In *Proceedings of CVPR*, volume 1, 4.

Litjens, G.; Kooi, T.; Bejnordi, B. E.; Setio, A. A. A.; Ciompi, F.; Ghafoorian, M.; van der Laak, J. A.; Van Ginneken, B.; and Sánchez, C. I. 2017. A survey on deep learning in medical image analysis. *Medical image analysis* 42:60–88.

Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of CVPR*, 3431–3440.

Novikov, A. A.; Lenis, D.; Major, D.; Hladuvka, J.; Wimmer, M.; and Bühler, K. 2018. Fully convolutional architectures for multi-class segmentation in chest radiographs. *IEEE Transactions on Medical Imaging*.

Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; and Lerer, A. 2017. Automatic differentiation in pytorch. In *NIPS-W*.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Proceedings of NIPS*, 91–99.

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

Sarker, M.; Kamal, M.; Rashwan, H. A.; Banu, S. F.; Saleh, A.; Singh, V. K.; Chowdhury, F. U.; Abdulwahab, S.; Romani, S.; Radeva, P.; et al. 2018. SLSDeep: Skin lesion segmentation based on dilated residual and pyramid pooling networks. In *Proceedings of International Conference on Medical Image Computing and Computer Assisted Intervention*.

Shiraishi, J.; Katsuragawa, S.; Ikezoe, J.; Matsumoto, T.; Kobayashi, T.; Komatsu, K.-i.; Matsui, M.; Fujita, H.; Kodera, Y.; and Doi, K. 2000. Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. *American Journal of Roentgenology* 174(1):71–74.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; and Fergus, R. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Tramèr, F.; Kurakin, A.; Papernot, N.; Goodfellow, I.; Boneh, D.; and McDaniel, P. 2018. Ensemble adversarial training: Attacks and defenses. In *Proceedings of ICLR*.

Van Ginneken, B.; Stegmann, M. B.; and Loog, M. 2006. Segmentation of anatomical structures in chest radiographs using supervised methods: a comparative study on a public database. *Medical image analysis* 10(1):19–40.

Wang, X.; Girshick, R.; Gupta, A.; and He, K. 2018. Non-local neural networks. In *Proceedings of CVPR*.

Xie, C.; Wang, J.; Zhang, Z.; Zhou, Y.; Xie, L.; and Yuille, A. 2017. Adversarial examples for semantic segmentation and object detection. In *Proceedings of ICCV*.

Yuan, Y. 2017. Automatic skin lesion segmentation with fully convolutional-deconvolutional networks. *arXiv preprint arXiv:1703.05165*.

Zhang, H.; Dana, K.; Shi, J.; Zhang, Z.; Wang, X.; Tyagi, A.; and Agrawal, A. 2018. Context encoding for semantic segmentation. In *Proceedings of CVPR*.