Defending Deep Learning-based Biomedical Image Segmentation from Adversarial Attacks: A Low-cost Frequency Refinement Approach

Qi Liu[1], Han Jiang[1], Tao Liu[2], Zihao Liu[2], Sicheng Li[3], Wujie Wen[1], and Yiyu Shi[4]

[1] Lehigh University, Bethlehem PA 18015, USA
{qil219,haj219,wuw219}@lehigh.edu
[2] Florida International University, Miami FL 33199, USA
{tliu023,zliu021}@fiu.edu
[3] Alibaba DAMO Academy, Sunnyvale CA, USA
sicheng.li@alibaba-inc.com
[4] University of Notre Dame, Notre Dame IN 46556, USA
yshi4@nd.edu

Abstract. Deep learning has demonstrated superb performance and efficiency in medical image segmentation. However, recently the community has also found the first practical adversarial example crafting algorithm dedicated to misleading deep learning-based biomedical image segmentation models. The generated segmentation-oriented adversarial examples, while almost indistinguishable by human eyes, can always produce target incorrect segmentation prediction with high intersection-over-union (IoU) rate, significantly concerning the safe use of such an emerging technique in medical diagnosis tasks. On the other hand, research on defending such an emerging attack in the context of medical image segmentation is lacking. In this work, we make the very first attempt to develop a low-cost and effective input-transformation based defense technique. To maximize the defense efficiency (or recovered segmentation results) of adversarial samples while minimizing the segmentation performance loss of benign samples after applying defense, we propose a novel low-cost image compression-based defense approach guided by fine-grained frequency refinement (FR). Extensive experimental results on various deep learning segmentation models show that our defense can offer very high defense efficiency against adversarial examples with very marginal segmentation performance loss of benign images on both ISIC skin lesion segmentation challenge and the problem of glaucoma optic disc segmentation. To further validate our method's effectiveness, we also extend our evaluation to the image classification model. We show the influence of our recovered segmentation prediction by our defense on disease prediction in adversarial settings. The code is released at: https://github.com/qiliu08/frequency-refinement-defense.

1   Introduction

Recently, deep learning has been widely applied to a variety of medical imaging tasks to assist doctors in making a more precise medical diagnosis at significantly reduced labor expenses. [13,4]. As the state-of-the-art deep learning solution continues advancing the performance in medical field, a question that naturally arises would be whether the adversarial examples could also compromise deep learning medical diagnosis results. The so-called adversarial example is a type of malicious inputs that possesses an indistinguishable visualization with benign inputs for human beings [8,3,11]. It has proved to be a significant security concern for deep learning-based image recognition and segmentation in computer vision. As expected, the latest study discovered the very first practical adversarial example in the context of medical image segmentation, namely Adaptive Segmentation Mask Attack (ASMA). It confirmed its high effectiveness to mislead the segmentation prediction to any target adversarial mask with a very high intersection-over-union (IoU) rate [14]. Distinct from existing segmentation attack algorithms for computer vision tasks, which are nontarget thus unrealistic in the medical field, ASMA is a targeted adversarial example generating algorithm that could lead to a convincing prediction shape of choice with subtle input modifications that are invisible to the bare eye. As a result, there is no doubt that such an emerging security flaw, if left unchecked, will cause severe consequences in the near future, especially considering that the doctors may ever-increasingly rely on deep learning predicted results for the treatment to patients. From an economic aspect, many fraud insurance claims could happen just because of the false breast cancel classification result incurred by these adversarial examples [7].

Rather than treating such an ever-increasing threat as an "after-thought" factor in deep learning-based medical imaging, in this paper, we make the very first attempt to mitigate such a new attack. Our approach is inspired by JPEG– a popular low-cost image compression framework [18,12]. We found that JPEG cannot balance the defense efficiency against adversarial examples and accuracy on benign images because of the human vision-centered quantization. To overcome this limitation, we propose the frequency refinement (FR) approach to redesign quantization based on the unique statistical pattern of adversarial perturbations in frequency domain. Our FR can almost recover the low IoU of segmentation prediction under adversarial settings to the original level without scarifying the accuracy of benign images (see the example on Fig. 2 column 1 and 4). We also extend our elevation to classification-based disease diagnosis. Experiments show that the false segmentation prediction incurred by adversarial examples would hurt the accuracy of disease classification significantly, e.g. increasing misdiagnosis risk (sensitivity: 50.6%). However, the recovered segmentation prediction by our FR can improve the sensitivity to the level comparable to the original model (68.2% vs. 71%). *To the best of our knowledge, this is the first study that targets the defense against the state-of-the-art adversarial examples in the context of medical imaging.*
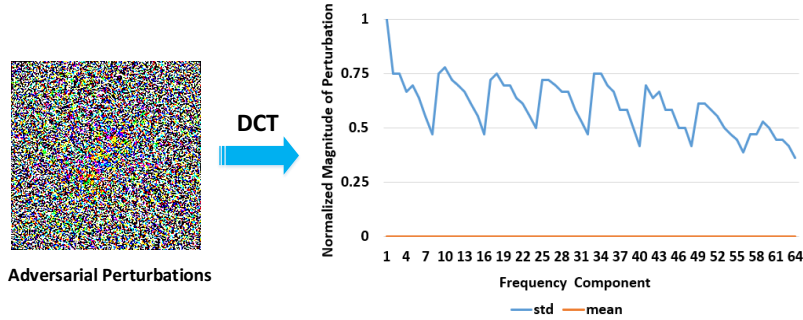
Fig. 1. Statistical information of adversarial perturbations in frequency domain

2 Background and Motivation

2.1 Adversarial Examples for Medical Image Segmentation

Ozbulak et al. [14] proposed a gradient descent-based targeted adversarial example generating algorithm, namely Adaptive Segmentation Mask Attack (ASMA). It is the first attack algorithm to accurately produce targeted adversarial masks meaningful in the sense of medical image segmentation. The basic idea of ASMA is to increase the prediction likelihood of the selected foreground pixels in the target adversarial mask for a specific class c except for pixels that are already predicted as class c by the segmentation model $f(\theta, X)$, while reducing the prediction likelihood of all other pixels outside the target adversarial mask for other classes except for those classified as other classes already. The input X can be iteratively updated as follows:

$$X_{i+1} = X_i + \alpha_i \cdot \sum_{c=1}^{m} \nabla_x \left(\mathbb{1}_{Y_{AE}=c} \odot \mathbb{1}_{\arg\max(f(\theta, X_i)) \neq c} \odot f_c(\theta, X_i) \right) \tag{1}$$

Where α_i is an adaptive perturbation multiplier and controlled by $\alpha_i = \beta \times IoU(Y_{AE}, Y_i) + \tau$, by considering the IoU score at the i-th iteration. β and τ are hyper-parameters to decide the final perturbation multiplier. $\odot$ is the element-wised matrix multiplication. Y_{AE} and X_0 are the desired prediction mask for adversarial example that contains class labels, and an input image, respectively.

2.2 JPEG Image Compression

JPEG [18] is a popular lossy image compression framework that mainly consists of image partitioning, Discrete Cosine Transform (DCT), quantization and lossless encoding etc. A 8×8 quantization table (Q-Table) with different quantization steps are used to scale DCT coefficients in frequency domain to achieve image compression. Based on the fact that human visual system (HVS) cares more about the low frequency features, larger (smaller) quantization steps are assigned to higher (lower) frequency components in Q-table. The trade-off between image quality and compression rate can be realized by scaling Q-table with a parameter–"Quantization Factor (QF)". A higher QF indicates better image quality but a lower compression rate.
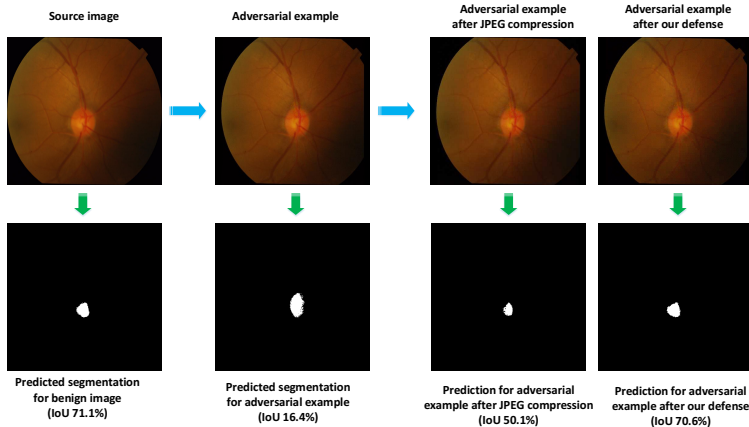
Fig. 2. The visualization of predicted segmentation for source and adversarial images with/without our defense and JPEG compression on the glaucoma optic disc dataset.

Limitation of JPEG on Defense. We select JPEG for our initial exploration of defense solutions against ASMA, considering its low-cost and popularity for medical image compression. Fig. 1 illustrates the ASMA-based adversarial perturbations of a representative image from glaucoma optic dis dataset [17], as well as the statistical information (mean and standard deviations) of the perturbations' DCT coefficients across 64 frequency bands. We observe that such adversarial perturbations could appear at any frequency band, with maximized distortions introduced in the lowest frequency band. However, JPEG compression always quantizes less on low frequency features (more on high frequency features) to preserve the visual quality for human eyes. Thus, standard JPEG can not effectively filter these perturbations in the whole spectral domain. As Fig. 2 shows, ASMA-based adversarial example drops IoU from 71.1% (benign image) to 16.4% on segmentation prediction (see the first two columns). However, JPEG-based defense only recovers IoU from 16.4% to 50.1% for adversarial example to maintain a high IoU for the benign image after defense given that such defense will be applied to any image in practice (see the third column). It confirms that standard JPEG is not an effective defense method against such adversarial examples and prompts the need to explore better low-cost defense solutions in the context of medical image segmentation.

3   Our Approach

In this section, we propose a frequency refinement (FR) approach to redesign the quantization table in JPEG compression, a.k.a. defensive quantization table (DQ-Table), to achieve both competitive testing accuracy and defense efficiency.

3.1   Defensive Quantization Analysis

To better explore the defensive quantization, we first analyze why quantization can mitigate adversarial examples but degrade the accuracy of benign images.

The DCT transformation is a linear function which transforms adversarial perturbations from spatial domain to spectral domain. The process can be as follows:

$$DCT(X + P) = DCT(X) + DCT(P) = C_X \cdot B + C_P \cdot B \tag{2}$$

Where P is the adversarial perturbation. C_X (C_P) and B are the DCT coefficients of input X (P) and DCT basis function, respectively. Here, $C_P \ll C_X$. The quantization function can be represented as:

$$q(X, QS) = \begin{cases} 0 & \text{if } \frac{|X|}{QS} < 0.5 \\ 1 & \text{otherwise} \end{cases} \tag{3}$$

where QS is the quantization step. As a result, the quantization process in spectral domain can be approximated as:

$$q(C_X + C_P, QS) \approx q(C_X, QS) + q(C_P, QS) \tag{4}$$

Ideally, if $QS > 2 \cdot |C_P|$, the malicious perturbation can be filtered. However, in practice, it is impossible to know the actual magnitude of the adversarial perturbation in different frequency bands. Instead, a possible solution is to use a large quantization step as much as possible. However, this can easily lead to prominent segmentation accuracy reduction due to introducing quantization errors to many important benign features $\sigma < \frac{QS}{2}$. Note a larger QS will increase the upper bound of quantization error, resulting in a more significant accuracy drop. To solve this issue, we propose to balance this by designing the defensive quantization table (DQ-Table) through fine-grained frequency analysis.

3.2 Frequency Component Analysis

To design an effective DQ-Table, we need to identify which frequency components are more critical for a deep learning-based segmentation model. For a single-pixel x, the 8×8 DCT can be expressed as:

$$x = \sum_{i=1}^{8} \sum_{j=1}^{8} c_{ij} \cdot b_{ij} \tag{5}$$

Where c_{ij} and b_{ij} are the DCT coefficient and DCT transformation basis at 64 frequency bands, respectively. The contribution of a frequency component b_{ij} to the segmentation model is associated with the gradient of the segmentation model function f with respect to b_{ij}, which can be calculated as:

$$\frac{\partial f}{\partial b_{ij}} = \frac{\partial f}{\partial x} \times \frac{\partial x}{\partial b_{ij}} = \frac{\partial f}{\partial x} \times c_{ij} \tag{6}$$

Eq. 6 means that the output of the segmentation model f will be mainly decided by the importance of pixel x–$\frac{\partial f}{\partial x}$ and the magnitude of DCT coefficient (c_{ij}). Since $\frac{\partial f}{\partial x}$ varies from one pixel to another, for each benign image, the importance of a frequency component for segmentation should be characterized from the statistics information of DCT coefficients. The study [15] has proved that the

DCT coefficients approximately obey a Laplace distribution with zero mean and various standard deviations (δ_{ij}). A larger δ_{ij} means a more critical contribution to the segmentation model. A simple solution to design the DQ-Table is to apply a larger (or smaller) QS at frequency bands with a larger (or smaller) δ_{ij}, so as to prevent the accuracy reduction induced by quantization errors. However, in our exploration, we also observe that a few frequency bands carrying the largest δ_{ij} suffer from more significant adversarial perturbation than others. This finding indicates that the frequency band with a larger δ_{ij}, which makes more essential contributions to the segmentation model, is also more prone to adversarial attacks. Thus, we further propose the frequency refinement approach (FR) for creating effective DQ-Table.

3.3   DQ-Table Design

Based on the above analysis, our proposed frequency refinement will guide the DQ-Table design based on the following two strategies: defense priority strategy and accuracy compensation strategy. According to two strategies, the 64 frequency bands will be roughly divided into three groups: priority defense (PD) band, accuracy compensation (AC) band and global defense (GD) band. Each group of bands will be assigned with an optimized quantization step (QS) constraint. We first sort the standard deviations (δ_{ij}). Then we set two thresholds T_1 and T_2 to categorize the 64 frequency bands into three bands. To simplify our design, we design a step function to assign QS constraint to the three divided bands based on the following strategies:

Defense Priority Strategy – We observe that adversarial perturbations are largely distributed in a few frequency bands with the first several largest δ_{ij}. For those frequency bands, we need to prioritize defense efficiency by setting a moderate QS_1, e.g. 20-50, to properly filter adversarial perturbations without distorting benign features. We name those frequency bands as priority defense (PD) band ($\delta_{ij} < T_1$). Besides, we find that the PD band usually consists of a few lowest frequency bands (see Fig. 1). This is because the test medical image sample usually carries adequate low-frequency information (e.g., a large number of similar pixels in the skin or blood image) compared with other types of image samples (e.g., ImageNet dataset [6]). Note that, the traditional JPEG compression will always set a smaller QS for lower frequency components, and hence is incapable of eliminating adversarial perturbations in these components, which is often the case for adversarial images targeting medical segmentation.

Accuracy Compensation Strategy – To compensate the accuracy reduction induced by quantization errors on PD band, we need to set a tiny QS_2, e.g. $QS_2 \leq 20$, in some of the sub-important frequency bands, namely accuracy compensation (AC) band ($T_1 < \delta_{ij} < T_2$). For other frequency bands, we adopt a large QS_3, e.g. 50-100, to eliminate perturbation as much as possible, namely global defense (GD) band ($\delta_{ij} > T_2$). As a result, the step function can be represented as:

$$f_s(\delta_{ij}) = \begin{cases} QS_1 & \text{if } \delta_{ij} < T_1 \\ QS_2 & \text{if } T_1 < \delta_{ij} < T_2 \\ QS_3 & \text{if } \delta_{ij} > T_2 \end{cases} \tag{7}$$

Table 1. Testing accuracy (IoU (%) on benign images) of JPEG v.s. our FRs.

	original model	JPEG(85)	JPEG(50)	FR(20,10,60)	FR(30,5,80)
ISIC skin lesion [5]	87.1±4.5	**86.5±7.5**	86±8.1	86.3±7.3	**86.5±7.2**
Glaucoma [17]	70.5±7.1	69.2±10.1	67.3±10	69.1±9.8	**69.3±9.6**

Table 2. Defense efficiency (IoU (%) on adversarial examples) of JPEG v.s. our FRs.

	no defense	JPEG(85)	JPEG(50)	FR(20,10,60)	FR(30,5,80)
ISIC skin lesion [5]	46.6±9.2	84.3±8.4	85.3±8.9	84.9±8.5	**85.5±8.3**
Glaucoma [17]	28.6±16.8	63.4±11.8	65.9±11	67.3±10.9	**68±10.1**

where $QS_1 \geq QS_2$ and $QS_2 \leq QS_3$. To simplify our design, we fix the number of frequency bands in each group band. Specifically, for the PD band, it includes three frequency bands with top-3 of sorted δ_{ij}. The AC band has 12 frequency bands (rank 4th - 15th of sorted δ_{ij}). The remaining frequency bands (rank 16th - 64th of sorted δ_{ij}) will be allocated to the GD band.

4   Evaluation

In this section, we comprehensively evaluate the defense efficiency (or accuracy recovery) against ASMA-based adversarial examples, as well as the segmentation accuracy of benign medical images for our proposed frequency refinement (FR) method. The advantage of segmentation prediction recovery on disease diagnosis (classification) using our defense is also discussed.

Datasets and Models – We used the two datasets in [14] to reproduce ASMA-based adversarial examples for a fair evaluation of our defense, which includes ISIC skin lesion segmentation dataset [5] and glaucoma optic disc dataset [17]. For segmentation task, two state-of-the-art segmentation models are selected, including Resnet-50 [2,9] for ISIC skin lesion dataset and U-Net [1,16] for glaucoma optic disc dataset. For disease diagnosis on glaucoma dataset, an ensemble of deep learning models [1], which consist of DenseNet [10] and ResNet[9], is adopted for disease classification purpose.

Evaluation Metrics – We characterize the intersection over union (IoU) of the label mask and the predicted segmentation mask on the given 100 benign and adversarial images. For the classification task, we evaluate the sensitivity of disease diagnosis. In our experiment, the disease diagnosis is a simple two-class classification problem: with or without the disease.

Evaluated Designs and Settings – To optimize DQ-Table of our FR approach, we designed multiple sets of (QS_1,QS_2,QS_3) to search for optimized results. Specifically, we set the following QS constraints (see Sec.3.3) in our optimization process: $QS_1 \in \{20, 30, 40, 50\}$, $QS_2 \in \{5, 10, 15, 20\}$, $QS_3 \in \{50, 60, 80, 100\}$. Among them, we selected two representative solutions – $FR(20, 10, 60)$ and $FR(30, 5, 80)$ as our defense candidates. We compare these two candidates with the JPEG compression approach with two representative QF configurations: 85 and 50, i.e., JPEG(85) and JPEG(50). JPEG(85) is a high quality baseline as it incurs a very marginal accuracy reduction for benign images, which is similar to our approach. JPEG(50) is a low quality baseline, which provides better defense efficiency at the cost of a more severe accuracy reduction.

Table 3. Defense efficiency (sensitivity (%)) of JPEG v.s. our FR for disease diagnosis.

	no defense	JPEG(85)	JPEG(50)	FR(20,10,60)	FR(30,5,80)	baseline
Glaucoma [17]	50.6%	54.6%	59.1%	65.5%	**68.2%**	71%

4.1   Evaluation Results

Testing Accuracy – As Table 1 shows, our two FR candidates show only marginal accuracy (IoU) degradation (i.e., $\sim 1\%$) on benign images for both ISIC and Glaucoma datasets. Due to the simplicity of ISIC dataset (i.e., a high IoU 87.1% on the original model), we also found that the low quality JPEG(50) compression can still achieve a competitive IoU, i.e., 86%. However, for a more complex Glaucoma dataset, JPEG(50) exhibits a prominent IoU degradation ($\sim 3\%$). Considering that defense will be applied to any benign or adversarial image in practice and the achievable defense efficiency makes more sense if the defense does not degrade the IoU of benign images, we focus on the JPEG(85) baseline for defense efficiency comparison.

Defense Efficiency – As listed in Table 2, our two FR candidates offer the most significant IoU improvement (or recover) on adversarial examples for both ISIC and Glaucoma datasets. In particular, FR(30,5,80) delivers better defense efficiency comparing with FR(20,10,60). This is because larger quantization steps (QS) used in the PD band and GD band can eliminate adversarial perturbations more effectively while smaller QS in the AC band can still compensate for the incurred accuracy reduction. Compared with JPEG baselines, our FRs achieve a slight improvement of defense efficiency for the simple ISIC dataset. However, for the more complex Glaucoma dataset, our two FRs significantly outperform the JPEG compression for the similar high IoU on benign images, e.g. 63.4% (JPEG(85)) vs. 67.3% (FR(20,10,60)) and 68% (FR(30,5,80)). Moreover, we also observe that even the more defensive JPEG(50) is still worse than any of the two FRs. These results indicate that our defense can better mitigate such adversarial examples with almost no accuracy loss, which is practically useful.

Further Evaluation on Disease Diagnosis – While ASMA-based segmentation adversarial examples do not directly target classification models, the generated false segmentation predictions can also impact the disease diagnosis accordingly. As Table 3 reports, the original disease classification sensitivity drops from 71% (basline) to 50.6% for images under adversarial settings on the segmentation task. We also observe that the IoU difference between our FR and JPEG on the segmentation task is further enlarged in disease diagnosis for defending against adversarial examples. For example, our FR(30,5,80) candidate surpasses JPEG(85) by 4.5% on segmentation adversarial examples, while such difference is increased to 13.6% (sensitivity) in classification. A possible explanation is that our defense, compared with JPEG, provides a more precise focus on extracting important features for classification. Therefore, our higher defense efficiency on segmentation can translate into better sensitivity on classification.

5   Conclusion

Deep learning has demonstrated tremendous performance improvement in fields like medical image segmentation. However, a recent study revealed that deep

learning-based segmentation models could be misled by the very first practical adversarial examples in the context of medical image segmentation. Our work is the first to explore and design defense solutions to mitigate such an emerging threat. According to the unique property of adversarial examples of medical images in spectral domain, we develop a frequency refinement (FR) approach to effectively eliminate malicious perturbations of adversarial examples without distorting benign features essential for segmentation. Experiments show that our FR can achieve high testing accuracy and defense efficiency simultaneously, serving as a reference design for future defense development in medical domain.

6   Acknowledgements

References

1. Agrawal, V., Kori, A., Alex, V., Krishnamurthi, G.: Enhanced optic disk and cup segmentation with glaucoma screening from fundus images using position encoded cnns. arXiv preprint arXiv:1809.05216 (2018)
2. Anand: isic 2018 github. `https://github.com/cygnus77/isic-2018` (2019)
3. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: 2017 ieee symposium on security and privacy (sp). pp. 39–57. IEEE (2017)
4. Chen, H., Dou, Q., Yu, L., Qin, J., Heng, P.A.: Voxresnet: Deep voxelwise residual networks for brain segmentation from 3d mr images. NeuroImage **170**, 446–455 (2018)
5. Codella, N.C., Gutman, D., Celebi, M.E., Helba, B., Marchetti, M.A., Dusza, S.W., Kalloo, A., Liopyris, K., Mishra, N., Kittler, H., et al.: 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: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). pp. 168–172. IEEE (2018)
6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
7. Finlayson, S.G., Chung, H.W., Kohane, I.S., Beam, A.L.: Adversarial attacks against medical deep learning systems. arXiv preprint arXiv:1804.05296 (2018)
8. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
10. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708 (2017)
11. Liu, Q., Liu, T., Liu, Z., Wang, Y., Jin, Y., Wen, W.: Security analysis and enhancement of model compressed deep learning systems under adversarial attacks. In: 2018 23rd Asia and South Pacific Design Automation Conference (ASP-DAC). pp. 721–726. IEEE (2018)

12. Liu, Z., Xu, X., Liu, T., Liu, Q., Wang, Y., Shi, Y., Wen, W., Huang, M., Yuan, H., Zhuang, J.: Machine vision guided 3d medical image compression for efficient transmission and accurate segmentation in the clouds. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 12687–12696 (2019)
13. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV). pp. 565–571. IEEE (2016)
14. Ozbulak, U., Van Messem, A., De Neve, W.: Impact of adversarial examples on deep learning models for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 300–308. Springer (2019)
15. Reininger, R., Gibson, J.: Distributions of the two-dimensional dct coefficients for images. IEEE Transactions on Communications 31(6), 835–839 (1983)
16. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
17. Sivaswamy, J., Krishnadas, S., Chakravarty, A., Joshi, G., Tabish, A.S., et al.: A comprehensive retinal image dataset for the assessment of glaucoma from the optic nerve head analysis. JSM Biomedical Imaging Data Papers 2(1), 1004 (2015)
18. Wallace, G.K.: The jpeg still picture compression standard. IEEE transactions on consumer electronics 38(1), xviii–xxxiv (1992)