FCSN: Global Context Aware Segmentation by Learning the Fourier Coefficients of Objects in Medical Images

Young Seok Jeon, Hongfei Yang, Mengling Feng

***Abstract*— The encoder-decoder model is a commonly used Deep Neural Network (DNN) model for medical image segmentation. Conventional encoder-decoder models make pixel-wise predictions focusing heavily on local patterns around the pixel. This makes it challenging to give segmentation that preserves the object's shape and topology, which often requires an understanding of the global context. In this work, we propose a Fourier Coefficient Segmentation Network (FCSN)—a novel global context-aware DNN model that segments an object by learning the complex Fourier coefficients of the object's masks. The Fourier coefficients are calculated by integrating over the whole contour. Therefore, for our model to make a precise estimation of the coefficients, the model is motivated to incorporate the global context of the object, leading to a more accurate segmentation of the object's shape. This global context awareness also makes our model robust to unseen local perturbations during inference, such as additive noise or motion blur that are prevalent in medical images. We compare FCSN with other state-of-the-art global context-aware models (UNet++, DeepLabV3+, UNETR) on 5 medical image segmentation tasks, of which 3 are camera imaging datasets (ISIC_2018, RIM_CUP, RIM_DISC) and 2 are medical imaging datasets (PROSTATE, FETAL). When FCSN is compared with UNETR, FCSN attains significantly lower Hausdorff scores with 19.14 (6%), 17.42 (6%), 9.16 (14%), 11.18 (22%), and 5.98 (6%) for ISIC_2018, RIM_CUP, RIM_DISC, PROSTATE, and FETAL tasks respectively. Moreover, FCSN is lightweight by discarding the decoder module, which incurs significant computational overhead. FCSN only requires 29.7M parameters which are 75.6M and 9.9M fewer parameters than UNETR and DeepLabV3+, respectively. FCSN attains inference and training speeds of 1.6ms/img and 6.3ms/img, which is 8× and 3× faster than UNet and UNETR. The code for FCSN is made publicly available at *https://github.com/nus-mornin-lab/FCSN*.**

***Index Terms*— Medical Image Segmentation, Global Context-aware Learning, Decoder-Free Segmentation.**

I. Introduction

Over recent years, we have witnessed increasing popularity in the applications of Deep Neural Network (DNN) for various medical image segmentation tasks. The encoder-decoder model [1], [2] is currently the most widely adopted DNN approach for the segmentation task. Given enough training data, the encoder-decoder models can extract local patterns from an image that are associated with labels at each spatial coordinate. However, due to its heavy reliance on local patterns, the model often fails to exploit the global contexts that potentially help to nullify nuisance local variations.

Specifically, in medical imaging tasks where the risk of misclassification is high, we need a robust model for many unpredictable local variations by incorporating global contexts. Taking the segmentation of optic cup in retinopathy as an example which is demonstrated in figure 1, the following problems are difficult to address unless the model learns the global context:

- anatomically, the shape of an optic cup is always like a single filled oval, but current DNN often gives segmentation with multiple components or with holes
- an optic disc has a smooth contour, but the current DNNs give contours with sharp corners or unnecessary zigzags
- retinopathy images from different sources are likely to suffer from different degradations, which cause generalization problems for current DNNs.

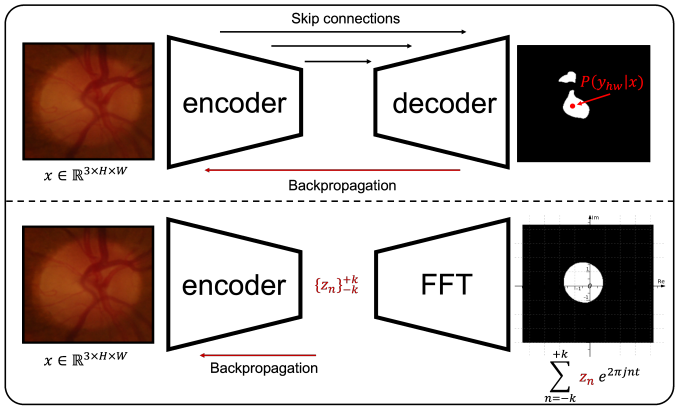


Fig. 1: Comparison of encoder-decoder model (**upper**) and Fourier Coefficient Segmentation Network (FCSN) (**lower**). Unlike the encoder-decoder model, which makes a coordinate-wise prediction of an object, our FCSN predicts the complex Fourier coefficients of the objects masks, which requires learning broader contextual information. Moreover, FCSN is more memory-efficient with the absence of a decoder.

In this paper, we argue that these problems, which are either ignored or indirectly treated in the conventional encoder-decoder segmentation models, can be effectively addressed if we train the DNN to directly predict the shape, size, and location of an object.

A. Encoder-decoder Segmentation Model

As shown in the first row of figure 1, modern segmentation models typically adopt an encoder-decoder structure which models a conditional probability of predicting label y_{hw} given an input $\mathbf{x}$ at each spatial coordinate h, w ($i.e.$ $\mathrm{p}(y_{hw}|\mathbf{x})$). The model is then optimized to maximize the likelihood of the spatially summed log probability ($i.e.$ $\mathrm{argmax}_{\mathrm{p}} \sum_{hw} y_{hw} \log \mathrm{p}(y_{hw}|\mathbf{x})$), assuming spatial independence across the coordinates. Based on the structure of the model and the way in which the model is optimized, the existing encoder-decoder model will make a prediction mainly relying on local patterns and often does not utilize the global context of the image at all. This absence of global context can cause inconsistency in segmentation performance, especially for the tasks that assume specific global priors.

Most of the existing works on global context learning aim to solve the problem by proposing a more flexible (general) model structure

Manuscript received xx xx, 2022; revised xx xx, xxxx , ... , and xx xx, xxxx; accepted xx xx, xxxx. Date of publication xx xx, xxxx; date of current version xx xx, xxxx.

This research is supported by the National Research Foundation Singapore under its AI Singapore Programme (Award Number: AISG-100E-2020-055 and AISG-GC-2019-001-2A) and the NMRC Health Service Research Grant (MOH-000030-00).

YS. Jeon, H. Yang, M. Feng are with the Saw Swee Hock School of Public Health and Institute of Data Science, National University of Singapore, Singapore (e-mail: youngseokjeon74@gmail.com; hfyang@nus.edu.sg; ephfm@nus.edu.sg).

Equal contribution: Young Seok Jeon and Hongfei Yang.

Corresponding author: Mengling Feng (ephfm@nus.edu.sg).

that offers the model an opportunity of capturing global patterns [3]–[5]. However, offering the opportunity does not necessarily mean that the model will explore the new aspect of learning. There is a possibility that the model will still focus on finding local shortcut evidence and hence fails to focus on the global evidences [6]. Also, higher flexibility could negatively impact the model performance when the network is trained under a data constraint. In this regard, we argue that increasing the model flexibility alone is an unstable solution to the global context learning problem.

B. Contribution

We propose a novel segmentation model—Fourier Coefficient Segmentation Network (FCSN) that lifts segmentation to a shape prediction task, representing the shape as Fourier coefficients. As shown in figure 1, FCSN perceives the segmentation mask as a smooth function in a complex domain, which can be accurately approximated as complex Fourier coefficients. We use Fourier Transform to extract the Complex Fourier coefficients of the contour of the mask. Hence, FCSN learns the global shape of an object by predicting its Fourier coefficients, and during inference, a contour is retrieved with Inverse Fourier Transform.

To motivate how predicting Fourier Coefficients helps to learn global context, imagine we want to segment an ellipse-shaped object, which can be precisely described by three complex Fourier Coefficients z_{-1}, z_0, z_1. The z_0 describes the center of the ellipse, and z_{-1} and z_1 determine the lengths and orientations of the semi-major and semi-minor axes. Thus, for a DNN to precisely predict the three coefficients, the model must learn to perceive the whole ellipse as a single object. This is in contrast to the traditional encoder-decoder model, where the model makes predictions only by looking at the local structure of the object.

Also, we propose to add a Fourier differentiable spatial to numerical transform (F-DSNT) module [7] to improve the accuracy of Fourier coefficient prediction and also to reduce memory consumption. One could view the coefficient prediction as a typical regression problem and introduce fully-connected (FC) layers on top of the spatially flattened feature. However, FC layers have several drawbacks: 1) they are over-parameterized, affecting the generalizability, 2) it assumes a fixed input shape; and 3) the output range is not bounded. Instead, DSNT drives the encoder module to produce heatmaps that represent the probability distributions of Fourier coefficients. DSNT does not introduce any trainable parameter and works with any input shape.

We evaluate the performance of FCSN on 5 Medical image segmentation tasks, which include skin lesion, optic disc, optic cup, prostate, and fetal head. FCSN outperforms state-of-the-art segmentation models such as DeepLab-v3+ and U-Net++ when evaluated with Hausdorff Distance. Furthermore, as our model can attend to global features, its performance does not degrade from local perturbations such as contrast change, additive noise, or motion blur. Lastly, our model is lightweight, requiring less computational cost by discarding the decoder module that has been indispensable in the modern segmentation model and incurs a considerable memory overhead.

II. Related Work

A. Encoder-decoder Models

FCN [8] and U-Net [1] were the early few DNN models that proposed encoder-decoder structure for semantic segmentation. However, the two approaches often produced noisy predictions that contained holes or non-smooth contours, implying that the models failed to understand the global context. The issue had been addressed broadly in two ways while preserving the encoder-decoder structure: by 1) increasing the receptive field size and 2) introducing a regularizer that penalizes non-smooth prediction.

1) Broader Receptive Field: For a unit in the prediction of a network, the theoretical receptive field (TRF) of this unit refers to the region in the input image that contributes to the prediction of this unit. For convolution neural networks, the TRF is usually only a fraction of the input image, which depends on the architecture and filter sizes of the networks. To make more global aware predictions, the TRF must be large enough to cover the whole region that contains information related to the prediction.

In the literature, several methods have been proposed to increase TRF. In [4], the authors proposed ParseNet, which incorporated a global context feature that is generated using a global pooling operation in feature embedding. In [2], the authors proposed DeepLab with Atrous Convolution module to increase TFR. Atrous convolution introduces extra spacing in the kernel, which provides a wider field of view with the same computational cost.

With recent advances in Transformer models [9], [10], which are sequential methods, people have been adopting Transformer structures to computer vision models to broaden their receptive field. In [11], [12], the authors proposed non-local U-Nets, which included Transformer modules [9] to extract long-range features, and in [13] the authors applied Transformer structure to medical image segmentation models.

As observed in [14], the effective receptive field (ERF) can be very different from the theoretical receptive field. The ERF is defined as the collection of pixels inside TRF that have a non-negligible impact on the prediction. It is found in [14] that for neural networks before training, the ERF is usually smaller than TRF, and proper training is needed to enlarge ERF. Therefore, models with large TRF may not be capable of effectively understanding the global context. In [15], the authors proposed the Lovász metric, which is a convex function that approximates the Intersection over Union (IoU) metric. Since IoU is calculated over the whole image, the proposed metric can facilitate global learning.

2) Regularizing Prediction: Another approach to promote smooth segmentations is to adopt regularization on the models or the predicted masks. In [16], the authors proposed the ACNN-Seg for predicting high-resolution segmentation masks from low-resolution images. They introduced an extra autoencoder (AE) network to regulate segmentation outputs, such that the AE would produce similar features for both the predicted masks and the ground-truths.

More recently, the authors in [17], [18] proposed to add spatial regularization to softmax activation functions to minimize the total variation of predictions, such that the predicted masks are more robust to various local perturbations in the images.

B. Segmentation via Shape

Most DNNs make per-pixel predictions for segmentation masks. One way to obtain more regularized prediction is to predict the shape of the segmentation mask, which effectively reduces the output dimensionality and complexity.

In [19], [20], the authors proposed DNNs that learn the parametrization of boundary curves via piecewise Bézier curves. However, the Bézier parametrization does not necessarily converge to the true boundary curve. In [21], the authors proposed to predict polar coordinates of sampled points on boundary curves for instance segmentation.

For classical image processing methods, the Fourier descriptor is a widely used technique to use Fourier transformation to encode boundaries of objects for image shape analysis and shape matching [22]–[25]. There are not many DNN approaches that utilize Fourier

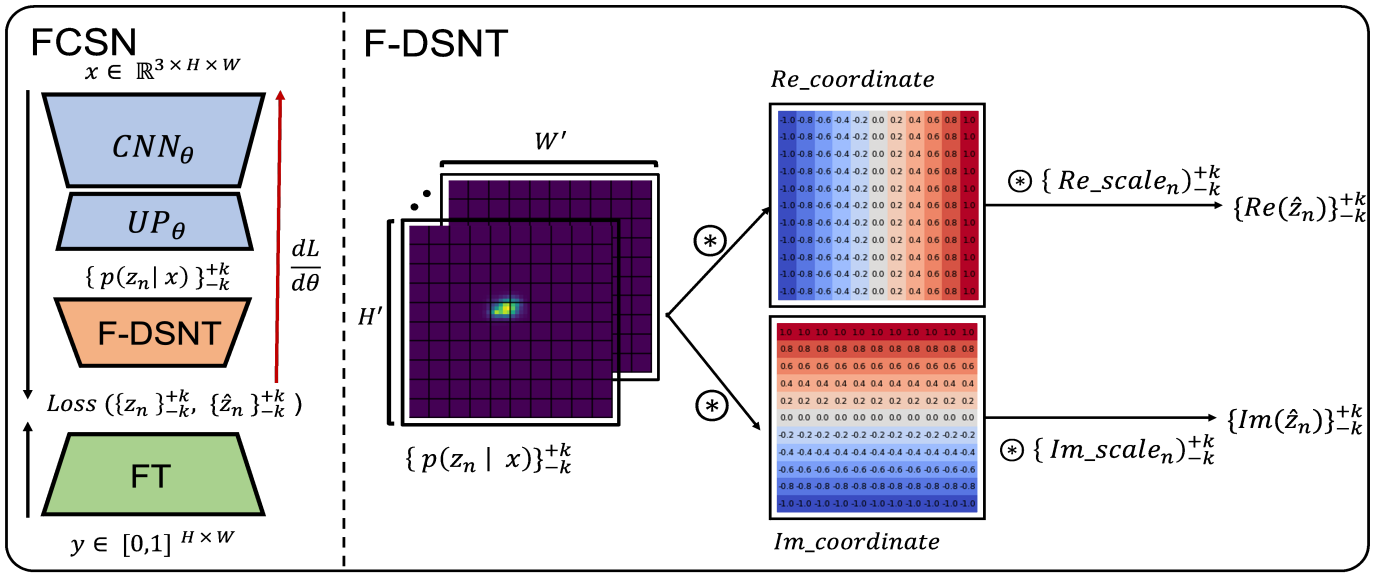


Fig. 2: Overview of the FCSN architecture (**left**) and differentiable spatial to numerical transform (DSNT) (**right**)

transforms for segmentation. In [26], the authors used DNN to learn Fourier coefficients of sampled points on boundary curves for instance segmentation. However, they regarded the x and y coordinates of boundary points as two sequences of real numbers and applied Fourier transforms independently. In our approach, we regard the boundary curve as a sequence in the complex domain, and we apply Complex Fourier Transform only once to get Fourier coefficients.

III. Proposed Method

As shown in figure 2, our DNN model consists of four modules. The first module CNN_θ is a feature extraction module that takes an image as its input. Any standard CNN backbone can be adopted. The second module UP_θ generates heatmaps which represent the discrete probability distribution functions (PDF) of Fourier coefficients. The third module F-DSNT "softly" picks up the most probable Fourier coefficient from each of the PDFs. The last module FT recovers segmentation masks from the predicted Fourier coefficients. To understand our approach, we first explain how we convert masks to Fourier coefficients. Also, the code for FCSN is made publicly available at *https://github.com/nus-mornin-lab/FCSN*.

*1) **FT** : Segmentation Masks to Fourier Coefficients:* Let Y be a binary segmentation mask. We regard Y as a function on the complex domain $D = \{x+jy : -1 \le x, y \le 1\}$, where $Y(x+jy) = 1$ for foreground and $Y(x+jy) = 0$ for background. Let $\alpha : [0,1] \to \mathbb{C}$ be a parametrization of the boundary curve of foreground. We assume α is a complex valued smooth curve with $\alpha(0) = \alpha(1)$. Given the boundary curve α, the region enclosed by α is the segmentation region.

The Fourier coefficients $\{z_n \in \mathbb{C}\}$ of the boundary curve $\alpha(t)$ is defined by

$$z_n = \int_0^1 \alpha(t) e^{-2\pi j n t}\, \mathrm{d}t \tag{1}$$

for $n = \ldots, -1, 0, 1, \ldots$, where j is the imaginary unit. The original boundary curve α can be fully recovered from the Fourier coefficients $\{z_n\}$ by taking the Inverse Fourier transform defined by

$$\alpha(t) = \sum_{n=-\infty}^{\infty} z_n e^{2\pi j n t}. \tag{2}$$

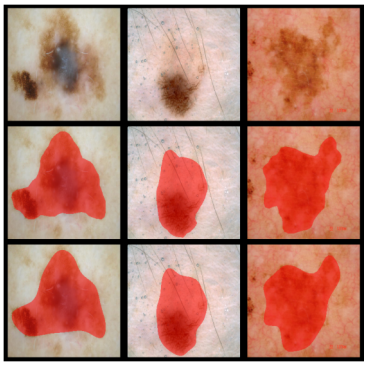


Fig. 3: Original mask (second row) and masks generated from boundary curves with 21 Fourier coefficients (third row).

Therefore, instead of making a direct prediction of the segmentation mask Y, it is possible to predict the Fourier coefficients $\{z_n\}$ and recover the mask Y with Inverse Fourier transform.

Predicting Fourier coefficients forces the training of DNN to utilize global context better. As suggested by equation (1), the Fourier coefficients, which we predict, are obtained by integrating global information on the boundary curve. This forces DNN models to learn the global context of an image better, facilitating to make more spatially consistent segmentation.

It is usually sufficient to only learn to predict the lower Fourier coefficients, which encode the location and the general shape of the boundary curve α. This is because the coefficients $\{z_n\}$ are concentrated on small absolute values of n when α is smooth: In fact, if α is k-times continuously differentiable, then z_n converges to 0 faster than $1/|n|^k$ for large n. Discarding higher Fourier coefficients can be regarded as a regularization that smooths ground-truth boundary curves. Figure 3 shows segmentation masks obtained by only taking z_n for $-10 \le n \le 10$.

2) UP_θ : Probability Distribution of Coefficients: Given a feature extracted from a raw input using a CNN module, UP_θ generates heatmaps that represent the discrete PDFs of possible Fourier coefficients. (*i.e.* $\{\text{p}(z_n|\mathbf{x})\}_{-k}^{+k} = \text{UP}_\theta \circ \text{CNN}_\theta$). UP_θ module consists of a 2D transposed convolution layer with $2*k+1$ kernels,

followed by a softmax activation across spatial axes. 2D transposed convolution layer projects input features to a higher spatial resolution; thus, the generated heatmaps are more granular. We apply softmax to normalize the heatmaps such that it is non-negative and sum to one.

3) $\mathbf{F-DSNT}$: Selecting the Most Probable Coefficients: Finding the most probable coefficient from each discrete PDF ($i.e.$ $\hat{z}_n = \text{argmax}\, \text{p}(z_n|\mathbf{x})$) is not differentiable. To make it differentiable, we adopt DSNT [7], which can be viewed as a soft-argmax operation. This is done by calculating the expectations of the PDFs. As shown in figure 2, the expectations are calculated by performing a weighted sum of discrete PDF with real and imaginary coordinate values.

For the original implementation of DSNT in [7], the PDFs are assumed to have spatial range $[-1, 1] \times [-1, 1]$. In our model, we multiply the output of our DSNT module with scaling constants estimated by checking the range of each Fourier coefficient from the training dataset. This is equivalent to increasing the resolution of PDFs for higher Fourier coefficients which are usually close to zero.

4) Loss Function: Our loss function is a combination of weighted L_1 and L_2 losses plus the Jensen-Shannon (JS) divergence regularization. Given a batch of M input images $\{\mathbf{x}^{(m)}\}$, our predicted coefficients $\{\hat{z}_n^{(m)} : -k \leq n \leq k\}$, and the ground truth Fourier coefficients $\{z_n^{(m)} : -k \leq n \leq k\}$, the loss function is

$$\text{Loss}(z_n, \hat{z}_n) = \frac{1}{M} \sum_{m,n} \Big\{ w_n \left(|\hat{z}_n^{(m)} - z_n^{(m)}| + |\hat{z}_n^{(m)} - z_n^{(m)}|_2^2 \right) + \text{JS}(p(\hat{z}_n^{(m)}|\mathbf{x}^{(m)}) || \mathcal{N}(\hat{z}_n^{(m)}, \sigma I_2)) \Big\}, \tag{3}$$

where $p(\hat{z}|\mathbf{x})$ is the PDF generated by our UP_θ module. The w_n's are weight constants that we introduce to promote the learning of higher Fourier coefficients which are much smaller than lower coefficients, defined as

$$w_n = \min \left\{ 1 + \frac{1}{\max_i |z_n^{(i)}| + \varepsilon}, 10 \right\}.$$

The $\text{JS}(p(\hat{z}|\mathbf{x})||\mathcal{N}(\hat{z}, \sigma I_2))$ is the JS divergence between the PDF $p(\hat{z}|\mathbf{x})$ and the bivariate normal PDF $\mathcal{N}(\hat{z}, \sigma I_2)$ with the same mean. The covariance σ of the bi-normal PDF is a hyperparameter. The JS regularization is minimized when the heatmap matches with the Gaussian distribution, thus making sure our heatmaps of Fourier coefficients are unimodal and concentrate nicely around the true locations of the Fourier coefficients.

IV. Experiments

A. Evaluation Metrics

(a) (b) (c) (d)

Fig. 4: (a) Ground truth, (b)–(d) three predictions with the same Dice value 0.9 but Hausdorff distances (smaller is better) 11.3, 26.9 and 93.3 respectively. Note that the star shape in (b) is smaller than that in (a).

Let Y be a segmentation mask, and let $\hat{Y}$ be a mask predicted by a DNN model. To measure model performance, we use both the Dice metric and the Hausdorff distance defined by

$$\text{H}(Y, \hat{Y}) = \max \left\{ \sup_{Y(y)=1} d(y, \hat{Y}), \sup_{\hat{Y}(y)=1} d(y, Y) \right\}, \tag{4}$$

where $d(y, Y)$ is the Euclidean distance from the point y to the target in Y, and $d(y, \hat{Y})$ is defined similarly. The smaller the Hausdorff distance is, the better the approximation of $\hat{Y}$ is to Y, and $\text{H}(Y, \hat{Y}) = 0$ means Y and $\hat{Y}$ coincides completely.

The Dice metric is widely used in evaluating segmentation models. However, the Dice metric is not sensitive to changes in the shape and topology of the masks. This is demonstrated in figure 4, where (a) is the ground truth, and (b)–(d) are three predictions with the same Dice value 0.9. However, it is clear that Figure 4(b) gives the best segmentation, while the shape of the segmentation in (c) is wrong, and the topology of the segmentation in (d) is wrong. On the other hand, the Hausdorff distance is more sensitive to changes in shape and topology, and it can successfully pick up the best segmentation.

B. Datasets

We test our methods on both camera imaging and medical imaging datasets.

1) Camera imaging dataset: We use two publicly available dataset: 1) ISIC-2018 [27] and 2) RIM-ONE-DL [28].

i) The ISIC-2018 dataset contains 2,594 and 100 dermoscopic images with ground truth segmentation for training and validation, respectively. The test dataset is not publicly available. Hence, following conventions of other papers using ISIC, we report the final evaluation results using 5-fold cross-validation on the training dataset.

ii) The RIM-ONE-DL dataset consists of 313 and 172 retinographies from normal and glaucoma patients. All images include a manual segmentation of the disc and cup that have been assessed by experts. The dataset contains 341 and 149 training and testing samples, respectively. As suggested by the dataset provider, we perform a simple train-test split evaluation.

2) Medical imaging dataset: We use two publicly available datasets: 1) PROSTATE [29] and 2) FETAL [30].

i) The Prostate dataset contains 48 3D volumes of MR images, and the target is to segment prostate central gland and peripheral zone. We report the final evaluation results using 5-fold cross-validation on this dataset.

ii) The Fetal dataset contains 2D ultrasound images of the standard plane of the fetal head, and the target is to segment the fetal head. There are 999 images in training set with segmentation masks and 335 test images without segmentation masks. We report the final evaluation results using 5-fold cross-validation on the training dataset.

C. Implementation Details

During training and inference, images are resized to have size 256×256. For data augmentations, we used ColorJitter, random crop, and random flip for the RIM dataset, and we replaced random crop by resizing and random crop for the ISIC dataset. For all our training, we trained for 500 epochs with a batch size of 8, and we used the Adam optimizer [31] with a learning rate of $3e^{-4}$ without weight decay.

To generate Fourier coefficients, we sampled 71 points on boundary curves and used FFT to get the Fourier Coefficients, where the model only learns 21 lower coefficients ($i.e.$ $\{z_n\}_{-10}^{+10}$). These numbers are hyper-parameters which we fixed for all experiments. See Appendix II for the effects of varying these hyper-parameters.

D. Results

1) Precise Shape Prediction: We compare the performance of FCSN with different backbone settings against state-of-the-art

TABLE I: Dice & Hausdorff comparison between FCSN and baseline encoder-decoder models on camera imaging datasets. The standard deviation (std) is computed from 5-fold results. The best result is in bold, and statistically worse performing results are in gray.

	ISIC_2018		RIM_CUP		RIM_DISC			
Models	**Haus±std**	**Dice±std**	**Haus**	**Dice**	**Haus**	**Dice**	**# Parameter (M)**	**# Flops (G)**
UNet	25.00 ± 1.00	0.89 ± 0.01	22.35	**0.78**	10.78	0.96	31.39	55.84
UNet++	24.06 ± 0.80	0.89 ± 0.01	22.25	0.77	11.79	0.96	36.63	138.16
DeepLabV3 (ResNet50)	20.79 ± 1.09	**0.90 ± 0.01**	21.69	0.77	10.92	0.96	39.63	40.99
DeepLabV3+ (ResNet50)	20.80 ± 0.76	**0.90 ± 0.01**	22.25	0.77	11.27	0.96	39.76	43.31
DeepLabV3+ (ResNet50 + Lovász)	20.44 ± 1.34	**0.90 ± 0.01**	20.60	**0.78**	10.63	0.96	39.76	43.31
UNETR (VIT-B-16)	20.05 ± 0.76	**0.90 ± 0.01**	18.98	**0.78**	10.95	0.96	105.32	32.14
FCSN (ResNet50)	20.21 ± 0.88	0.88 ± 0.01	18.15	0.77	9.85	0.96	29.71	**23.54**
FCSN (DResNet26)	20.14 ± 1.00	0.88 ± 0.01	18.07	0.77	9.59	0.96	**22.16**	83.01
FCSN (DResNet50)	**19.14 ± 0.86**	0.88 ± 0.01	**17.42**	**0.78**	**9.16**	0.96	29.71	98.11

TABLE II: Dice & Hausdorff comparison between FCSN and baseline encoder-decoder models on medical imaging datasets. The standard deviation (std) is computed from 5-fold results. The best result is in bold.

	PROSTATE		FETAL	
Models	**Haus±std**	**Dice±std**	**Haus±std**	**Dice±std**
UNet	12.70 ± 1.93	0.80 ± 0.03	7.63±1.02	**0.97 ± 0.00**
UNet++	14.73 ± 2.14	0.81 ± 0.03	9.31±1.78	**0.97 ± 0.00**
DeepLabV3 (ResNet50)	11.25 ± 1.26	**0.82 ± 0.02**	6.33±0.52	**0.97 ± 0.00**
DeepLabV3+ (ResNet50)	12.18 ± 0.49	0.81 ± 0.03	6.04±0.57	**0.97 ± 0.00**
DeepLabV3+ (ResNet50 + Lovász)	11.75 ± 0.83	**0.82 ± 0.02**	6.01±0.57	**0.97 ± 0.00**
UNETR (VIT-B-16)	14.12 ± 0.49	0.81 ± 0.02	6.19±0.34	**0.97 ± 0.00**
FCSN (ResNet50)	11.23 ± 1.58	0.80 ± 0.02	6.58±0.37	0.96±0.00
FCSN (DResNet26)	**11.07 ± 1.20**	0.81 ± 0.01	**5.84 ± 0.26**	0.96±0.00
FCSN (DResNet50)	11.18 ± 0.92	0.81 ± 0.02	5.98±0.31	0.96±0.00

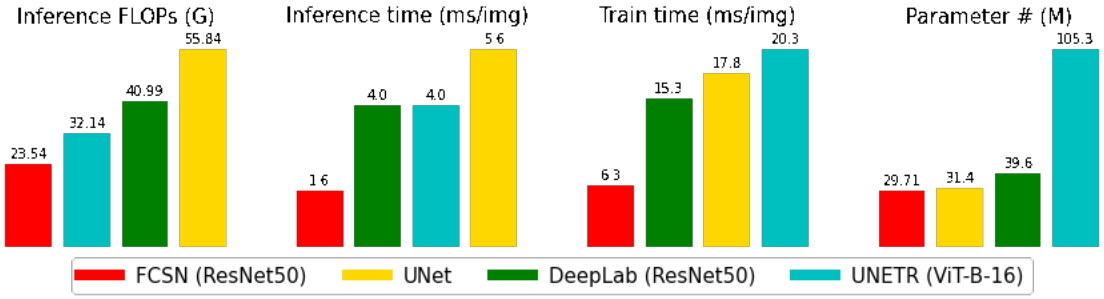


Fig. 5: Computational efficiency of FCNS and baseline models in 4 different aspects: Floating point operations (FLOPs), inference & train speed (ms/img), and model size (M).

segmentation models, including vanilla UNet [1], UNet++ [32] and DeepLab-v3+ [2] (with ResNet50 as its backbone) with/without the lovász-softmax loss [15], and UNETR [13] (with VIT-B-16 as its backbone). We perform experiments on 2 categories of medical images: camera imaging dataset (ISIC skin lesion, RIM_CUP, and RIM_DISC) and medical imaging dataset (PROSTATE and FETAL). The model performance is accessed with Hausdorff and Dice metrics.

As shown in Tables I and II, for all instances, FCSN achieves a lower Hausdorff score while maintaining a competitive Dice score, supporting that the shape of generated mask closely matches with ground truth. We note that the performance of FCSN improves when we use DResNet [33] backbone that produces higher resolution output. Also, using the deeper DResNet50 backbone for the camera imaging dataset further improves the performance. However, the DResNet26 backbone achieves the best performance for the medical imaging dataset.

Based on paired T-tests and results for all tasks, our FCSN method with DResNet26 or DResNet50 backbone outperforms all baseline methods with a significant level 0.05. We have provided full p-value matrices for test statistics in Appendix III.

2) Robustness to Perturbations: We test the robustness of models to four types of perturbations at inference: Gaussian noise, Salt & Pepper noise, contrast changes, and motion blur. All the models are not re-trained with perturbed data: they are all trained only with original data, which does not include any of the perturbation cases we test on. We chose Gaussian and Salt & Pepper noises because they are the most common additive and impulsive noises, respectively. Contrast change and motion blur are typical degradations in medical images. The results are summarized in figure 7, where the level of perturbation increases along the x-axis. Compared with the DeepLab-v3+ (with Lovász loss) and the UNETR models, our method is more robust, especially for the two noises, where our method can give almost consistent predictions regardless of noise level; on the other hand, the predictions of the DeepLab-v3+ model deteriorate

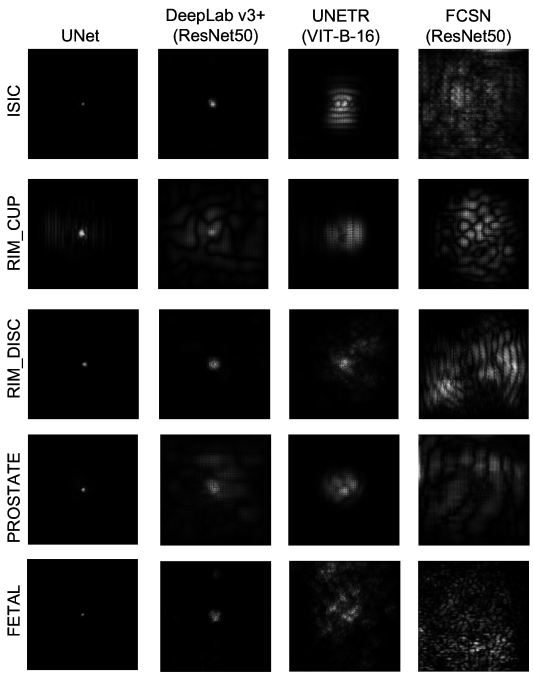


Fig. 6: Comparison of Effective Receptive Field (ERF)

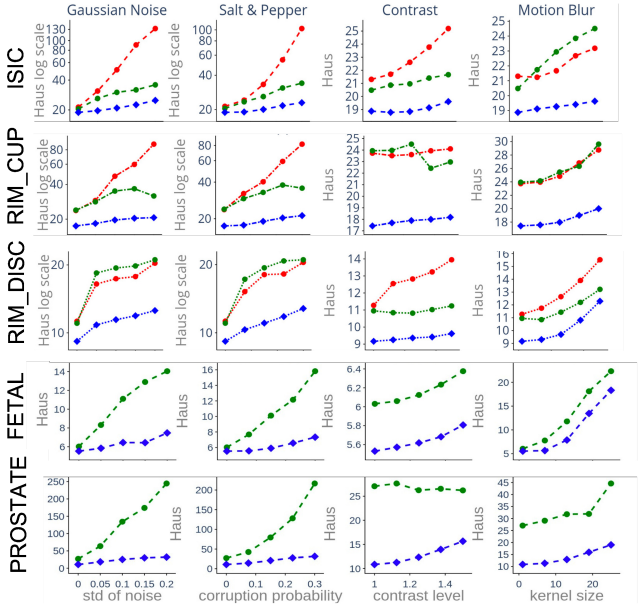


Fig. 7: Hausdorff distance (smaller the better) of inferences with perturbations. Red: DeepLab-v3+ with ResNet50 backbone (Lovász loss), Green: UNETR with VIT-B-16 backbone, Blue: FCSN with ResNet50 backbone. FCSN is more robust to perturbations, especially for heavy noises. The results of DeepLab-v3+ on PROSTATE and FETAL tasks are omitted because it frequently gives empty or near-empty prediction even when light noise is present, as shown in figure 9.

heavily as noise level increases. Metrics of results of the UNETR model are either similar to that of DeepLab-v3+ or lie between the DeepLab-v3+ and our method.

Figure 8 shows examples of segmentation results for images with perturbations (more examples in Appendix IV). For images with noise or contrast change, the DeepLab-v3+ method omitted large portions of target areas, and the UNETR failed to correctly segment the RIM cup with Salt & Pepper noise, while our method consistently gives reasonable segmentation for all cases. For the image with motion blur, the DeepLab-v3+ and UNETR methods wrongly included a large portion of the background area. All the predictions of the DeepLab-v3+ have either the wrong shape or the wrong topology. On the other hand, our method gives satisfactory segmentation results.

3) Global Context Awareness: Here, we empirically prove that the two major strengths of FCSN, precise shape prediction and robustness to perturbations, indeed arise from the model's global context awareness. We propose to use the Effective Receptive Field (ERF), initially proposed by Luo et al. [14], as the method to measure the global context awareness of models. ERF measures how much each input pixel contributes to the model prediction. Mathematically, this is done by computing the partial derivative of an arbitrary output unit y_i with respect to input tensor $\mathbf{x}$ (i.e $\partial y_i / \partial \mathbf{x}$), measuring how much y_i changes as $\mathbf{x}$ changes by a small amount. ERF is therefore a natural measure of the importance of $\mathbf{x}$ with respect to y_i.

Figure 6 shows the comparison of ERF for various models. We observe that FCSN visually attains a significantly bigger ERF size compared to baseline models across all tasks, strongly supporting our global context awareness argument.

4) Computational Efficiency: We compare the computational efficiency of FCSN against baseline segmentation models. Specifically, we measure models' Floating point operations (FLOPs), inference time (ms/img), training time (ms/img), and parameter number (M). We compare FCSN with ResNet50 backbone against vanilla UNet, DeeLab with ResNet50 backbone, and UNETR with VIT-B-16 backbone. During the measure of FLOPs, inference & training time, we set the input size to 256 $\times$ 256. The results in Figure 5 show the computational efficiency of FCSN in all of the 4 aspects. Comparing with the least performing model for each of the aspects, FCSN requires 58% less FLOPs, 8$\times$ faster training and inference speed, and 5$\times$ less parameter number. Note that the computation overheads from Fourier and inverse Fourier transforms are small, which are equivalent to two 1D convolution layers with kernel size of 21 and input size 21. Empirically, these two transforms only take 0.05ms/img.

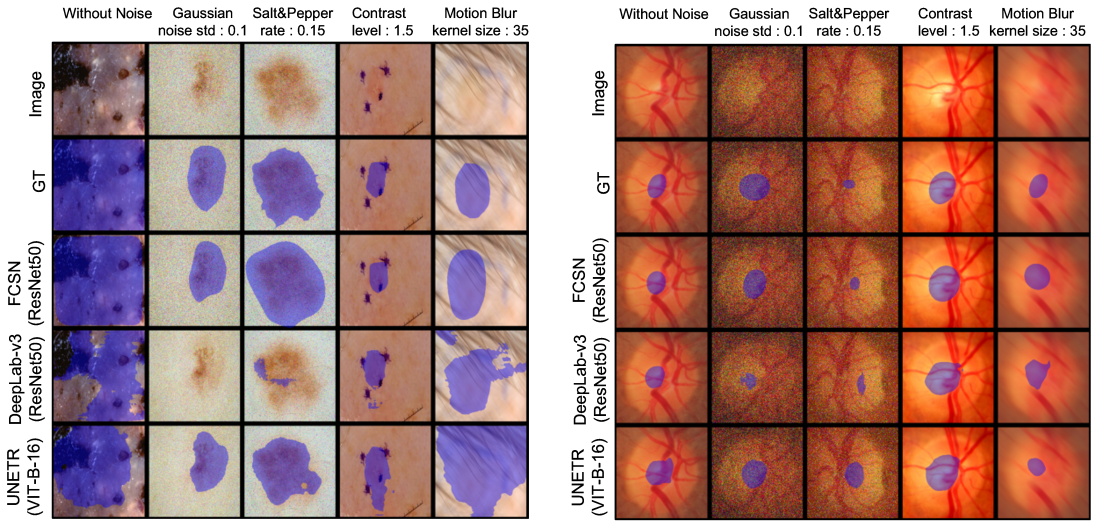
Our model has high computational efficiency because our model does not contain a conventional decoder. For most segmentation models employing neural network approach, they contain decoders that have several layers of 2D convolution and up-sampling operations. This will introduce a large number of model parameters and heavy computations. On the other hand, our model only contains the encoder, and the prediction of Fourier coefficients is based on the F-DSNT layer, which incurs little computation and does not contain learnable parameters.

E. Ablation

1) Impact of DSNT: For comparison, we remove $\mathbf{UP}_\theta$ and $\mathbf{DSNT}$ parts of our model and connect the feature maps from our backbone to FC layers to get Fourier coefficients. Experiment results in table III show that for the Dice metric, the DSNT approach consistently gives better results, while for the Hausdorff metric, the DSNT approach gives better results in most of the cases.

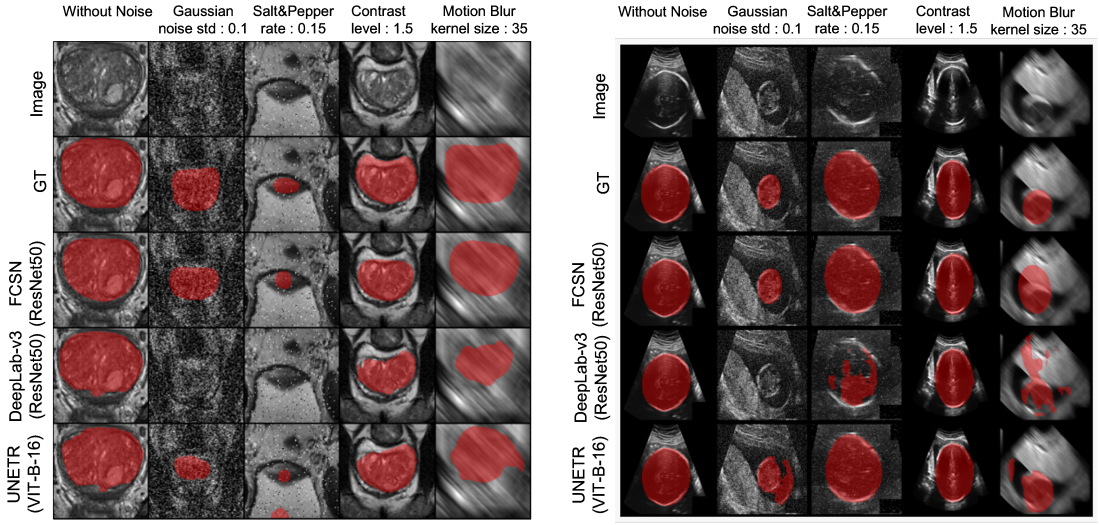
We argue that this is because the FC layers contain a significant number of learnable parameters, which made the training more difficult. On the other hand, DSNT method does not introduce extra learnable parameters. Our observation here is in consistence with findings in key point detection tasks [7], where DSNT approach has better performance than directly using FC layers.

2) Impact of JS Divergence: We study the effect of the Jensen-Shannon divergence regularization on our model by removing the regularization or by altering σ in the covariance σI_2 of the 2D Gaussian PDF. As seen from table IV, the introduction of the



(a) (b)

Fig. 8: Visual comparison of predicted masks (a) ISIC (b) RIM_CUP tasks with perturbations.



(a) (b)

Fig. 9: Visual comparison of predicted masks with perturbations (a) PROSTATE (b) FETAL.

regularization greatly improves model performance, but our model is not sensitive to the choice of σ.

We believe this is because the JS divergence can promote learning of unimodal probability density functions (PDF) regardless of the variance, which can be regarded as a regularization of the PDF. Regularizing using divergences is a common technique in PDF learning/estimation which improves model accuracy [34], [35].

V. Limitation and Future Works

There are a couple of future research directions that can make the proposed FCSN more robust.

1) 3D shape learning: MRI and CT scans are 3D in nature. To apply the current FCSN structure to 3D tasks, the scan must be interpreted as independent slices. However, the independent assumption across the slices could lead to an inconsistent mask prediction. As a solution to this, one can generalize our framework by modifying our 2D F-DSNT module to a 3D version of it.

2) High Variance of Higher Frequency Coeficients: Figure 10a shows that FCSN can give accurate predictions for the (-1)-th Fourier coefficients for ISIC task. The violin plots in Figure 10b show that the relative errors of the predicted Fourier coefficients become larger as the index of coefficients increase from -1 to -10 and from 1 to 10, and the (-1)-th and 0-th coefficients have the smallest errors. Note that the Python package we used produced clockwise boundaries. The (-1)-th coefficients correspond to clockwise circles that match the overall size of the clockwise boundaries and the 0-th coefficients correspond to the centers of the objects, which are the most prominent geometric features of masks in our setup. We propose the following conjectures for the larger errors of higher frequency coefficients:

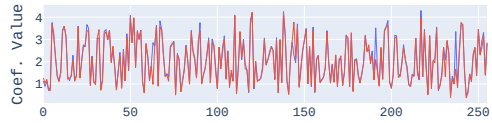
- Our current backbones have excessive pooling layers which reduce spatial resolution of output feature maps. Thus some information in higher frequencies may have been lost.
- Image masks can be noisy, which leads to larger noise in ground truth of higher frequency coefficients. This makes learning them difficult.

TABLE III: Dice and Hausdorff of FCSN (ResNet50) with DSNT or FC head.

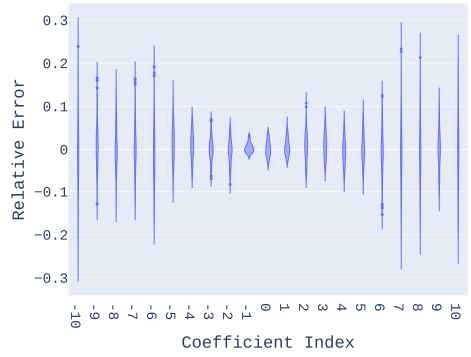
Tasks	Metric	Heads	Epoch Number 100	200	300	400
ISIC	Dice	DSNT	**0.87**	**0.88**	**0.89**	**0.89**
		FC	0.86	0.87	0.87	0.88
	Haus	DSNT	**21.64**	20.07	**20.11**	**19.82**
		FC	21.70	**19.91**	21.03	20.31
RIM_CUP	Dice	DSNT	0.74	0.76	**0.77**	**0.77**
		FC	0.74	0.76	0.76	0.76
	Haus	DSNT	19.16	**18.62**	**18.39**	**18.47**
		FC	**18.46**	19.04	18.59	19.09
RIM_DISC	Dice	DSNT	0.95	0.95	**0.96**	**0.96**
		FC	0.95	0.95	0.95	0.95
	Haus	DSNT	**10.04**	**9.45**	**9.42**	**9.46**
		FC	10.48	10.23	10.56	10.16

TABLE IV: Dice of FCSN (ResNet50) on ISIC task with varying regularisation.

	$\sigma = 0.005$	$\sigma = 0.01$	$\sigma = 0.015$	no JS
Dice	0.89	0.88	0.89	0.84



(a) Plot of real parts of (-1)-th Fourier coefficients for a batch of 256 images. Blue lines: Predictions by FCSN (ResNet50). Red lines: Ground truth.



(b) Violin plot of prediction errors (normalized by maximum value) of real parts of all Fourier coefficients.

Fig. 10: Error of predicted Fourier coefficients on ISIC validation set.

- Higher coefficients usually have much lower scales, which may hinder gradient flows when using stochastic gradient descent methods to optimize model parameters.

We leave proper investigation to future research.

3) Multi-object Segmentation Task: To extend FCSN to multi-instance segmentation cases such as multi-organ segmentation, one could fuse FCSN with MaskRCNN [36]. The MaskRCNN method performs multi-object segmentation in two steps: in the first step, for each object, the method predicts a bounding box that covers the whole object; in the second step, the method extracts the image patch inside the bounding box and perform a per-pixel segmentation prediction within the patch. We propose to replace the per-pixel segmentation step in MaskRCNN with our FCSN method.

4) Learning other transforms: FCSN learns to predict Fourier Coefficients for segmentation, and it works well for targets with smooth boundaries. However, if the target boundaries contain sharp corners, one may consider modifying FCSN to learn coefficients from more general transforms, like wavelet or tight frame transform. The idea is to use a proper family of base functions that are more efficient in coding boundary curves.

VI. Closing Remarks

In this paper, we propose FCSN, a novel and lightweight segmentation model that segments an object by predicting the Fourier coefficient of the object's contour. Our model is designed to incorporate the global context of an image, leading to more accurate segmentation that better preserves the shape and topology of the object. Moreover, global context awareness makes our model robust to unseen local perturbations during inference.

Our approach is the first step towards a systematic study of performing segmentation by predicting coefficients of mask decomposition. There are many other approaches besides predicting Fourier coefficients. For instance, one can use wavelet or tight frame transforms to obtain more efficient decomposition for boundary curves with sharp corners.

Appendix I
Pseudocode for Training with FCSN (ResNet50)

Algorithm 1 FCSN (ResNet50) training pseudocode

Require: Training images $\{x\}$ and corresponding Fourier coefficient vectors $\{\mathbf{z}\}$.
  $i \leftarrow 0, B \leftarrow 8$ ▷ Batch size 8
  net.encoder $\leftarrow$ *ResNet50.layers[:-2]*
  net.up-sampling $\leftarrow$ **UP** in Subsection III-.2
  net.DSNT $\leftarrow$ F-DSNT in Subsection III-.3
  $\theta \leftarrow$ net.weights
  adam $\leftarrow$ *Adam optimizer with parameter list* θ
  while $i <$ total epoch **do**
    while sample (without replacement) B images x **do**
      get feature maps: $y =$ net.encoder(x)
      up-sample feature maps: $\tilde{y} =$ net.up-sampling(y)
      get PDFs: $p =$ soft-max$(\tilde{y})$
      predict Fourier coefficients: $\hat{z} =$ net.DSNT(p)
      get loss: Loss$(z, \hat{z}) = L_1(z, \hat{z}) + L_2(z, \hat{z})$
      gradients $\leftarrow$ backward propagate Loss$(z, \hat{z})$
      $\theta \leftarrow$ adam(gradients)
    end while
  end while

Full code at *https://github.com/nus-mornin-lab/FCSN*.

Appendix II
How many Fourier coefficients shall we learn?

For all our experiments, FSCN is fixed to predict 21 Fourier coefficients (10 positive-order coefficients, 10 negative-order coefficients, and also the 0-th order coefficients), where the ground truth coefficients are 21 coefficients truncated from lower frequency parts of Fourier coefficients calculated from 71 sampling points on the boundary of segmentation mask. Note that using 21 lower Frequency coefficients calculated from 71 sampling points is different from sampling 21 points and use all their Fourier coefficients, where the

Fig. 11: Difference between using 21 Fourier coefficients from 21 sampling points and using 21 lower Frequency coefficients from 71 sampling points. Yellow pixels for false negative, and red pixels for false positive.

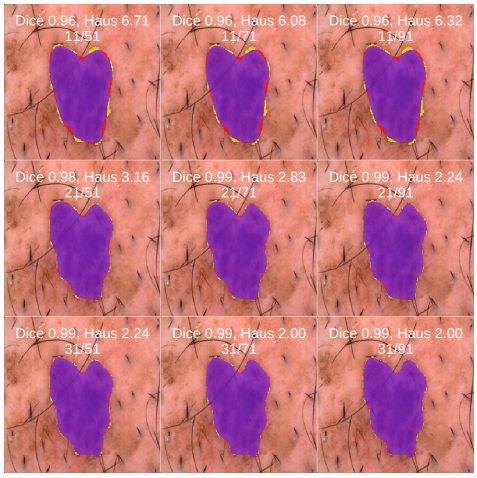


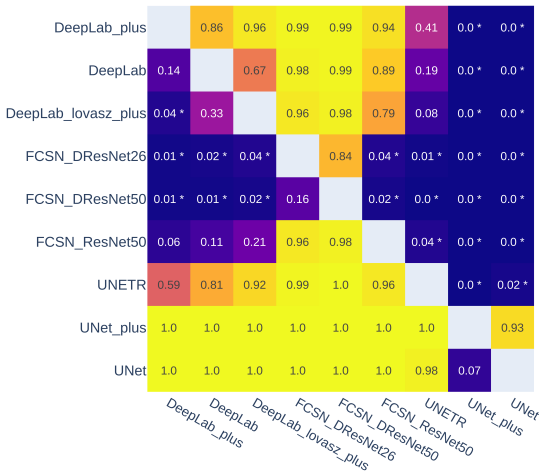
Fig. 12: Masks generated with varying number of sampling points and number of ground-truth Fourier coefficients. Yellow pixels for false negative, and red pixels for false positive.

latter may cause an aliasing problem, see figure 11. In general, one needs a high number of sampling points, but using only the first few lower frequency Fourier coefficients achieves a good recovery of the mask.

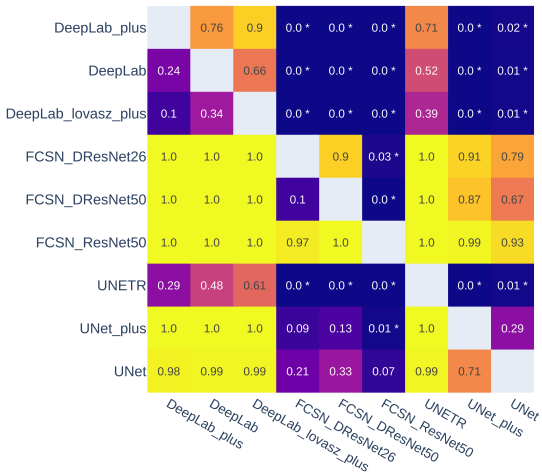
Figure 12 demonstrates masks generated with varying numbers of sampling points and number of ground-truth Fourier coefficients. We see that with our current setup, the mask generated from 21 lower frequency Fourier coefficients is virtually indistinguishable from the original mask (with Dice over 0.99). On the other hand, 11 Fourier coefficients always produce over-smoothed masks. Configurations with more sampling points or number of coefficients can produce slightly better masks, but they will lead to heavier data pre-processing load or slower training/inference time. Moreover, our experiments suggest that these slight improvements in recovering masks do not lead to better validation/testing results.

For the ISIC and Fetal dataset, we have done experiments to train FSCN to predict 11 Fourier coefficients, and we observe that for both dataset the accuracy of predictions are similar to FSCN with 21 Fourier coefficients with less than 1% difference. For the Fetal dataset, this is expected since the segmentation masks are always oval shaped, where 11 Fourier coefficients are usually enough to recover the ground-truth mask. For the ISIC data, as shown in Figure 12, 11 ground-truth Fourier coefficients produce over-smoothed masks with Hausdorff distance around 6. We argue that our observation from the experiment is because the segmentation task in ISIC is difficult to learn. From table I we see that for comparing methods which all predict per-pixel segmentation, the best Hausdorff distance is above 20, which is much higher than 6. Thus we conjecture that the over-smoothed label produced by 11 Fourier coefficient is not the bottleneck in training our FSCN on ISIC images.

We stress that the number of Fourier coefficients to learn for our



(a) p-value matrix for Hausdorff metric.



(b) p-value matrix for Dice metric.

Fig. 13: Statistical tests (significant level 0.05) for model performances on all dataset for the Hausdorff metric and the Dice metric. Value at (i, j) is p-value for testing the Hausdorff/Dice for the method at row i than the method at column j.

FSCN is a hyper-parameter, and users can always make plots like those in Figure 12 to pick up a good empirical value if our proposed 21 does not work for them.

APPENDIX III
STATISTICAL TESTS FOR NUMERICAL EXPERIMENTS

Figure 13 gives p-value matrices for testing different models for the Hausdorff and the Dice metrics. With DResNet26/DResNet50 backbones, our FCSN method consistently outperforms all the baseline methods for the Hausdorff metric with significant level 0.05.

APPENDIX IV
SEGMENTATION VISUALIZATION

We give segmentation results on clean images in figure 14a and 14b, where some bad segmentation parts are highlighted by yellow circles. We also visualize some segmentation results in figures 14c to 14f, where in each subfigure a single image is perturbed with various noises.

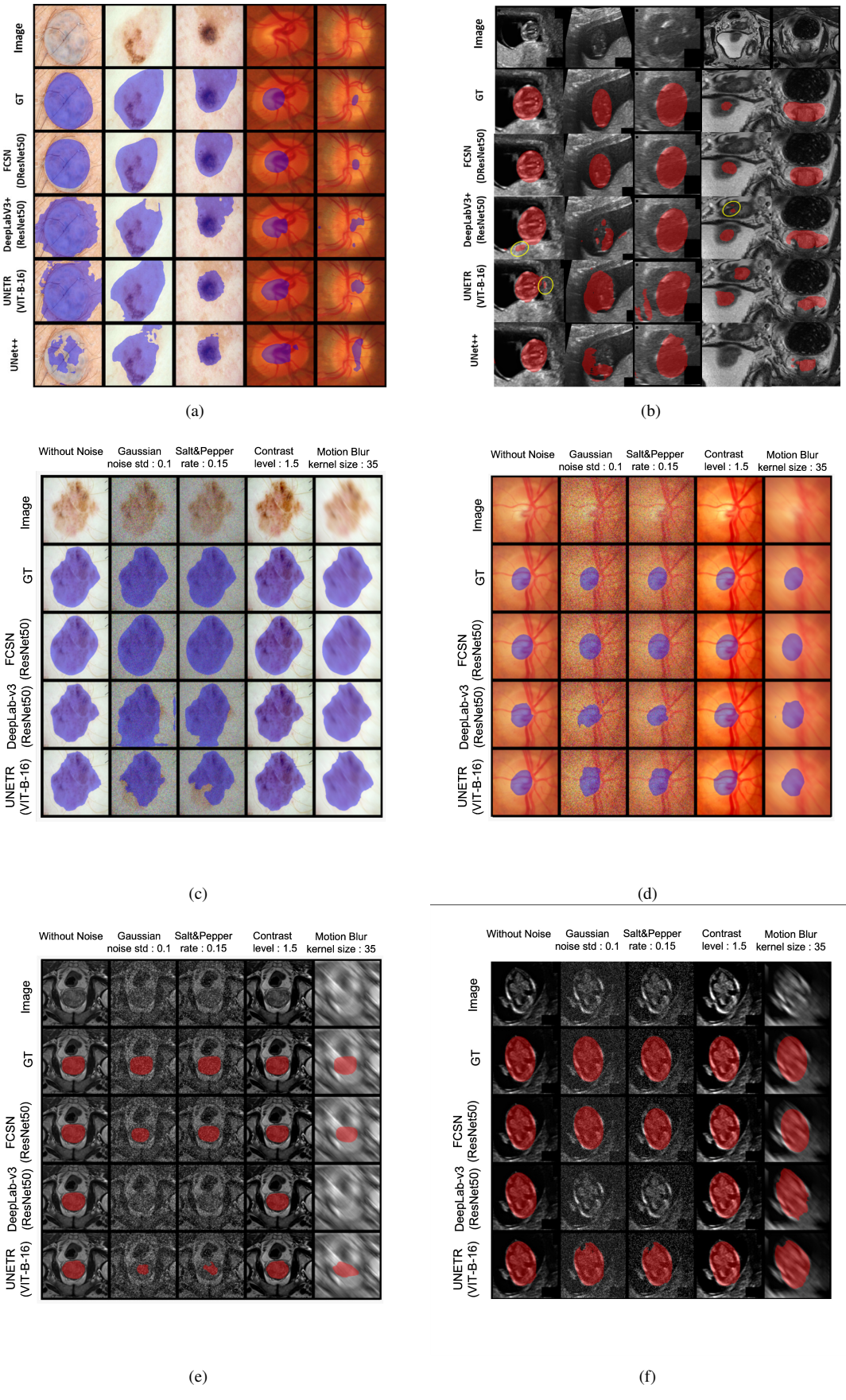


Fig. 14: Visualization of segmentation results on clean images (a) ISIC and RIM_CUP (b) FETAL and PROSTATE. Some small bad segmentation parts are highlighted by yellow circles. Visual comparison of predicted masks (c) ISIC (d) RIM_CUP tasks with perturbations on a single image. Visual comparison of predicted masks with perturbations (e) PROSTATE (f) FETAL on a single image.

REFERENCES

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

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

[3] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[4] W. Liu, A. Rabinovich, and A. C. Berg, "Parsenet: Looking wider to see better," *arXiv preprint arXiv:1506.04579*, 2015.

[5] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning hierarchical features for scene labeling," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1915–1929, 2012.

[6] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann, "Shortcut learning in deep neural networks," *Nature Machine Intelligence*, vol. 2, no. 11, pp. 665–673, 2020.

[7] A. Nibali, Z. He, S. Morgan, and L. Prendergast, "Numerical coordinate regression with convolutional neural networks," *arXiv preprint arXiv:1801.07372*, 2018.

[8] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[11] Z. Wang, N. Zou, D. Shen, and S. Ji, "Non-local u-nets for biomedical image segmentation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, 2020, pp. 6315–6322.

[12] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

[13] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu, "Unetr: Transformers for 3d medical image segmentation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 574–584.

[14] W. Luo, Y. Li, R. Urtasun, and R. Zemel, "Understanding the effective receptive field in deep convolutional neural networks," *Advances in neural information processing systems*, vol. 29, 2016.

[15] M. Berman, A. R. Triki, and M. B. Blaschko, "The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4413–4421.

[16] O. Oktay, E. Ferrante, K. Kamnitsas, M. Heinrich, W. Bai, J. Caballero, S. A. Cook, A. De Marvao, T. Dawes, D. P. ORegan *et al.*, "Anatomically constrained neural networks (acnns): application to cardiac image enhancement and segmentation," *IEEE transactions on medical imaging*, vol. 37, no. 2, pp. 384–395, 2017.

[17] F. Jia, J. Liu, and X.-C. Tai, "A regularized convolutional neural network for semantic image segmentation," *Analysis and Applications*, vol. 19, no. 01, pp. 147–165, 2021.

[18] F. Jia, X.-C. Tai, and J. Liu, "Nonlocal regularized cnn for image segmentation," *Inverse Problems & Imaging*, vol. 14, no. 5, p. 891, 2020.

[19] Y. Liu, H. Chen, C. Shen, T. He, L. Jin, and L. Wang, "Abcnet: Real-time scene text spotting with adaptive bezier-curve network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9809–9818.

[20] H. Chen, Y. Deng, B. Li, Z. Li, H. Chen, B. Jing, and C. Li, "Bezierseg: Parametric shape representation for fast object segmentation in medical images," *arXiv preprint arXiv:2108.00760*, 2021.

[21] E. Xie, P. Sun, X. Song, W. Wang, X. Liu, D. Liang, C. Shen, and P. Luo, "Polarmask: Single shot instance segmentation with polar representation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 12 193–12 202.

[22] D. Zhang and G. Lu, "Shape-based image retrieval using generic fourier descriptor," *Signal Processing: Image Communication*, vol. 17, no. 10, pp. 825–848, 2002.

[23] E. T. Bowman, K. Soga, and W. Drummond, "Particle shape characterisation using fourier descriptor analysis," *Geotechnique*, vol. 51, no. 6, pp. 545–554, 2001.

[24] Y. Rui, A. C. She, and T. S. Huang, "A modified fourier descriptor for shape matching in mars," in *Image databases and multi-media search*. World Scientific, 1997, pp. 165–177.

[25] I. Kunttu, L. Lepisto, J. Rauhamaa, and A. Visa, "Multiscale fourier descriptor for shape classification," in *12th International Conference on Image Analysis and Processing, 2003. Proceedings.* IEEE, 2003, pp. 536–541.

[26] H. U. M. Riaz, N. Benbarka, and A. Zell, "Fouriernet: Compact mask representation for instance segmentation using differentiable shape decoders," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 7833–7840.

[27] N. Codella, V. Rotemberg, P. Tschandl, M. E. Celebi, S. Dusza, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. Marchetti *et al.*, "Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic)," *arXiv preprint arXiv:1902.03368*, 2019.

[28] F. J. F. Batista, T. Diaz-Aleman, J. Sigut, S. Alayon, R. Arnay, and D. Angel-Pereira, "Rim-one dl: A unified retinal image database for assessing glaucoma using deep learning," *Image Analysis & Stereology*, vol. 39, no. 3, pp. 161–167, 2020. [Online]. Available: https://www.ias-iss.org/ojs/IAS/article/view/2346

[29] M. Antonelli, A. Reinke, S. Bakas, K. Farahani, A. Kopp-Schneider, B. A. Landman, G. Litjens, B. Menze, O. Ronneberger, R. M. Summers *et al.*, "The medical segmentation decathlon," *Nature communications*, vol. 13, no. 1, pp. 1–13, 2022.

[30] T. L. van den Heuvel, D. de Bruijn, C. L. de Korte, and B. v. Ginneken, "Automated measurement of fetal head circumference using 2d ultrasound images," *PloS one*, vol. 13, no. 8, p. e0200412, 2018.

[31] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

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

[33] F. Yu, V. Koltun, and T. Funkhouser, "Dilated residual networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 472–480.

[34] D. Yu, K. Yao, H. Su, G. Li, and F. Seide, "Kl-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7893–7897.

[35] K. Su and Z. Lu, "Divergence-regularized multi-agent actor-critic," in *International Conference on Machine Learning*. PMLR, 2022, pp. 20 580–20 603.

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