I^{2}Net: Exploiting Misaligned Contexts Orthogonally with Implicit-Parameterized Implicit Functions for Medical Image Segmentation

Jiahao Yu, Fan Duan, and Li Chen(✉)

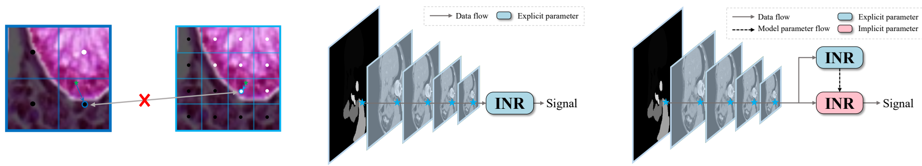
School of Software, BNRist, Tsinghua University, Beijing, China
{yujh21,df22}@mails.tsinghua.edu.cn, chenlee@tsinghua.edu.cn

Abstract. Recent medical image segmentation methods have started to apply implicit neural representation (INR) to segmentation networks to learn continuous data representations. Though effective, they suffer from inferior performance. In this paper, we delve into the inferiority and discover that the underlying reason behind it is the indiscriminate treatment for context fusion that fails to properly exploit misaligned contexts. Therefore, we propose a novel **I**mplicit-parameterized **I**NR **Net**work (**I^{2}Net**), which dynamically generates the model parameters of INRs to adapt to different misaligned contexts. We further propose novel gate shaping and learner orthogonalization to induce I^{2}Net to handle misaligned contexts in an orthogonal way. We conduct extensive experiments on two medical datasets, i.e. Glas and Synapse, and a generic dataset, i.e. Cityscapes, to show the superiority of our I^{2}Net. Code: https://github.com/ChineseYjh/I2Net.

Keywords: Medical image segmentation · Implicit neural representation · Feature misalignment · Dynamic neural network.

1 Introduction

Segmentation is a fundamental task in medical image analysis. Recent works [15,11] apply implicit neural representation (INR) to build decoders of segmentation networks for learning continuous data representations to tackle the drawback of conventional discrete grid-based data representations. These INR-based decoders model the segmentation map as a continuous signal field, which extracts a set of latent codes from the multi-scale feature maps for each continuous input coordinate and feeds them into a neural network, typically an MLP, to output the signal (Fig. 1b). Nevertheless, all the existing INR-based decoders bring about feature misalignment phenomenon, i.e. context mismatch among the extracted multi-scale latent codes caused by naive interpolation, e.g. nearest neighbor (Fig. 1a). Existing studies on feature misalignment argue that the context mismatch is harmful and *directly* results in the inferior performance of segmentation models, thus they design fancy aligning mechanisms in the decoder of segmentation



(a) Feature misalignment (b) Naive INR-based (c) Our **I^{2}Net**

Fig. 1. a) An example of context mismatch occurring at the input coordinate ($\star$) in INR. b) Naive INR-based decoders representing segmentation maps with explicit/static parameters. c) Our I^{2}Net representing maps with implicit/dynamic parameters, which is further represented with an explicit/static INR.

models to extract contextually matched features for each grid on the feature maps to improve performance [17,12,27].

However, we argue that the context misalignment is not always harmful to model performance. On the contrary, properly exploiting the misaligned contextual latent code is often beneficial for category discrimination in medical images. For example, a detected context of a large right liver can be exploited as evidence for raising the probability of detecting a gallbladder or a right kidney, even if the contexts of the latter are not detected. This suggests that a *heterogeneous* contextual code can also indirectly provide valuable inference evidence for the category discrimination of target coordinates, which we refer to as the *implicit discrimination patterns* of the context features. From this perspective, we argue that the underlying *direct* reason for the inferior performance of INR-based decoders comes from the **indiscriminate treatment** for contextual latent code fusion, which makes it difficult for the INRs to learn various implicit discrimination patterns. Specifically, existing INR-based decoders all adopt a low-capacity static MLP expecting to aggregate *homogeneous* contextual codes for input coordinates without considering their inherent difference [15], thus the fitting scope of their learned static model parameters hardly includes the various implicit discrimination patterns of contexts, leading to the unexpected misclassification when INRs encounter heterogeneous contextual codes at the input coordinates.

Therefore, to address the problem, this paper proposes a novel **I**mplicit-parameterized **I**mplicit neural representation **Net**work (**I^{2}Net**), to better exploit misaligned context latent codes by capturing their implicit discrimination patterns. Specifically, we first propose a high-capacity implicit-parameterized implicit function with the idea of dynamic networks [10], which dynamically generates the model parameters of the INR-based decoders based on the context codes (Fig. 1c), thereby adapting the INRs to various implicit discrimination patterns of contexts. The dynamic parameters are composed by weighting several shared parameter sets with the implicit gates modeled by another vanilla INR, where each shared parameter set (named as pattern learner, abbr. PL) is responsible for learning a discrimination pattern and the gates aim to perform soft selection on these learned patterns. Then, to induce PLs to capture different discrimination patterns, we further propose novel gate shaping and learner orthogonalization to achieve orthogonal exploitation, which introduces constraints from the view of implicit gates and PLs respectively. On one hand, gate shaping induces gate distribution to be sharp or smooth by controlling its entropy, thereby preventing network degeneracy including learning similar gates for PLs and the "rich get richer" phenomenon. On the other hand, our learner orthogonalization restricts the orthogonality among the gradients of segmentation loss over PLs to encourage PLs to learn orthogonal patterns.

To summarize, the major contributions are as follows: **1)** Different from the prior, we discover the underlying *direct* reason for the inferior performance of INR-based decoders for medical image segmentation, i.e. indiscriminate context fusion, and propose a novel method, I^2Net, to address the problem. **2)** For the first time, we propose a novel implicit-parameterized INR to adapt to various discrimination patterns of contexts, which is generic and can be directly applied to other INR-related areas. **3)** We further propose novel gate shaping and learner orthogonalization to induce PLs in our I^2Net to capture orthogonal discrimination patterns. **4)** We conduct extensive experiments on two medical datasets of different modalities, Glas [23] and Synapse [16], to demonstrate the superiority of our I^2Net. We further generalize our I^2Net to the generic semantic segmentation and conduct experiments on Cityscapes [5] to exhibit its superiority.

2 Method: I²Net

2.1 Preliminary

We present our method using 2D cases. Given a medical image $I \in \mathbb{R}^{C_I \times H \times W}$, the medical image segmentation task aims to predict a segmentation map $P \in \mathbb{R}^{C_P \times H \times W}$, where H, W and C_I are the height, width, and channel of the input image I, and C_P denotes the number of target classes. Typical INR-based methods [11,15] represent each segmentation map P with multi-scale feature maps extracted from an encoder, i.e. $\{F_i\}_{i=1}^N$ (N is the number of scale levels), and a shared static decoding function, which is defined to produce the signal map P. Given a continuous coordinate $\mathbf{p} \in \mathbb{R}^2$, the signal value is defined as

$$\mathrm{INR}(\mathbf{p}, \mathbf{F}; \mathbf{\Theta}) = f_{\mathbf{\Theta}}\Big(\big\{z_i^*, \mathbf{p} - p_i^*\big\}_{i=1}^N\Big) \tag{1}$$

where $\mathbf{F}$ denotes multiscale features $\{F_i\}_{i=1}^N$, $f_{\mathbf{\Theta}}$ is an MLP parameterized by $\mathbf{\Theta}$, z_i^* denotes the extracted latent code from $\mathbf{p}$ on the i-th feature map, and p_i^* is the coordinate of z_i^*.

2.2 Implicit-Parameterized Implicit Function

In contrast to the static-parameterized INR defined in Eq. (1), we propose an implicit-parameterized INR, whose parameters are dynamically generated by another INR to adapt to various discrimination patterns of contexts. Inspired by [10], we compose the dynamic parameters by weighting several shared parameter

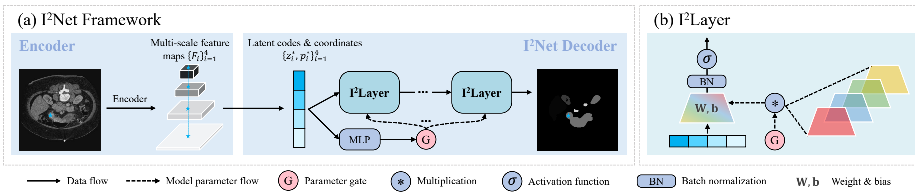


Fig. 2. The framework of I^{2}Net and the detailed design of I^{2}Layer (with $K = 4$).

sets with the dynamic gates modeled by an INR, hence the signal value at the coordinate $\mathbf{p}$ is defined as

$$\begin{aligned} \mathrm{I}^2\mathrm{Net}(\mathbf{p}, \mathbf{F}) &= \mathrm{INR}\big(\mathbf{p}, \mathbf{F}; \sum_{i=1}^{K} g_i(\mathbf{p}, \mathbf{F}; \mathbf{\Theta}_g) \times \hat{\mathbf{\Theta}}_i\big), \\ g(\mathbf{p}, \mathbf{F}; \mathbf{\Theta}_g) &= \sigma\big(\mathrm{INR}(\mathbf{p}, \mathbf{F}; \mathbf{\Theta}_g)\big) \end{aligned} \tag{2}$$

where K is the number of shared parameter sets, $\hat{\mathbf{\Theta}}_i$ is the i-th shared parameter set (named as pattern learner, abbr. PL), $g(\cdot; \mathbf{\Theta}_g)$ outputs a gate vector with a dimension of K and is parameterized by $\mathbf{\Theta}_g$, and $\sigma(\cdot)$ is the softmax function. In implementation, as shown in Fig. 2, our I^{2}Net is built by stacking multiple I^{2}Layers, each of which is a dynamic fully connected layer whose weight and bias are dynamically generated by

$$\left[\mathbf{W}_g^{(l)}, \mathbf{b}_g^{(l)}\right] = \sum_{i=1}^{K} g_i \left[\hat{\mathbf{W}}_i^{(l)}, \hat{\mathbf{b}}_i^{(l)}\right] \tag{3}$$

where $\mathbf{W}_g^{(l)}$ and $\mathbf{b}_g^{(l)}$ are dynamic weight and bias in the l-th I^{2}Layer, g_i is the i-th component of the gate g, $\hat{\mathbf{W}}_i^{(l)}$ and $\hat{\mathbf{b}}_i^{(l)}$ are static weight and bias in the l-th I^{2}Layer of the i-th PL ($\hat{\mathbf{\Theta}}_i$).

2.3 Orthogonal Exploitation of Contexts

Our I^{2}Net provides high model capacity to include various discrimination patterns, but we still need to introduce additional constraints for inducing I^{2}Net to capture those patterns. Since the dynamic parameters are generated with implicit gates and PLs, we design constraint losses from these two perspectives, i.e. gate shaping loss and learner orthogonalization.

Gate Shaping. We first empirically observe that directly training I^{2}Net is prone to degenerate solutions where the gating function tends to learn similar weights for all PLs. As a remedy, we first propose *gate instance sharpening loss*:

$$\mathcal{L}_{gis} = \frac{1}{B \times N_p} \sum_{i=1}^{B} \sum_{j=1}^{N_p} H\big(g(\mathbf{p}_j, \mathbf{F}_i; \hat{\mathbf{\Theta}}_g)\big) \tag{4}$$

where B is the batch size, N_p is the number of coordinate points sampled on each image during training, and $H(\cdot)$ is the entropy function, i.e. $H(p) = -\sum_k p_k \ln(p_k)$. $\mathcal{L}_{gis}$ induces gate distribution to be sharp by reducing the entropy of the gate of each coordinate point instance $\mathbf{p}_j$, thus preventing the degeneracy. However, solely utilizing $\mathcal{L}_{gis}$ leads to another degeneracy, i.e. "rich get richer" phenomenon, where one of the PLs is always picked and others ignored. Hence we further propose *gate expectation smoothing loss*:

$$\mathcal{L}_{ges} = -H\big(\frac{1}{B \times N_p} \sum_{i=1}^{B} \sum_{j=1}^{N_p} g(\mathbf{p}_j, \mathbf{F}_i; \hat{\boldsymbol{\Theta}}_g)\big) \tag{5}$$

which prevents "rich get richer" degeneracy by increasing the entropy of the expectation of the gate of the coordinate point instance $\mathbf{p}_j$. In Eq. (5), we use the average of the gates of all the sampled points in a batch to approximate the expectation. With these two losses, I^{2}Net is encouraged to assign high weights to different PLs when handling different discrimination patterns. Thus, the overall gate shaping loss is defined as $\mathcal{L}_{gs} = \mathcal{L}_{gis} + \lambda_{ges}\mathcal{L}_{ges}$, where λ_{ges} is a hyperparameter.

Learner orthogonalization. To induce each PL to specialize in learning distinct discrimination patterns, we apply gradient-based orthogonal regularization. The intuition behind this is that moving locally along the direction of the gradient leads to the biggest change in model prediction, while moving orthogonal to the gradient leads to the least change. Thus, we restrict the gradient of the segmentation loss over each PL to be orthogonal to each other to induce different PLs to focus on learning different patterns. Specifically, we first define the unit vector of the gradient over the i-th PL as $\boldsymbol{\nabla}_i = \mathrm{norm}\big(\mathrm{flat}(\frac{\partial \mathcal{L}_{seg}}{\partial \hat{\boldsymbol{\Theta}}_i})\big)$, where $\mathrm{flat}(\cdot)$ is flattening operation, $\mathcal{L}_{seg}$ is segmentation loss, and $\mathrm{norm}(\mathbf{x}) = \frac{\mathbf{x}}{\|\mathbf{x}\|}$ ($\|\cdot\|$ is Euclidean norm). Then our learner orthogonalization is defined as

$$\mathcal{L}_{lo} = \frac{1}{K(K-1)} \sum_{1 \le i < j \le K} |\boldsymbol{\nabla}_i^{\mathsf{T}} \boldsymbol{\nabla}_j|^2 \tag{6}$$

In implementation, $\mathcal{L}_{lo}$ is applied to weights and biases in the parallel PLs in a layer-wise manner, thus the loss is defined as $\mathcal{L}_{lo} = \sum_{l=1}^{N_L} \mathcal{L}_{lo}^{(l)}$, where N_L is the number of I^{2}Layers. Finally, the total loss for training I^{2}Net is defined as $\mathcal{L} = \mathcal{L}_{seg} + \lambda_{gs}\mathcal{L}_{gs} + \lambda_{lo}\mathcal{L}_{lo}$, where λ_{gs} and λ_{lo} are hyperparameters.

3 Experiments

3.1 Experimental Settings

Datasets. a) Glas [23] is a colon histology image dataset for binary gland segmentation. It provides 165 images of 512×512 resolution, which are split into

85 images for training and 80 for testing. b) Synapse [16] is a clinical CT image dataset for multi-organ segmentation, which contains 30 contrast-enhanced CT scans in 8 abdominal organs with 3779 axial CT images of 512×512 resolution in total. We follow [3] to use the split of 18 training cases (2212 axial slices) and 12 cases for validation. c) Cityscapes [5] is a popular urban scene dataset for generic semantic segmentation, which contains 19 classes and 5000 finely annotated images of 1024×2048 resolution, which are further split into 2975, 500, and 1525 images for training, validation, and testing respectively.

Evaluation metrics. For the medical datasets, we employ average dice score (DSC) and average 95% Hausdorff distance (HD95) to evaluate model performance. For Cityscapes, we adopt Intersection over Union averaged over classes (mIoU) for evaluation. The number of float-point operations (FLOPs) and the number of parameters (#Params) are also employed for efficiency evaluation.

Implementation details. We conduct experiments on one single NVIDIA RTX 3090 GPU for Glas and Synapse, and four for Cityscapes. We follow ConTrans [18], TransUNet [3], and IFA [11] to configure loss function, optimizer, learning rate scheduler, batch size, crop size, and training epochs (or iterations) for Glas, Synapse, and Cityscapes, respectively. We follow [11] to sample points during training, thus $N_p = \frac{H}{4} \times \frac{W}{4}$. We set λ_{gs}, λ_{ges}, and λ_{lo} to 0.25, 0.5, and 0.25, respectively. For Glas and Synapse, I^{2}Net has two I^{2}Layers with a hidden dimension of 128, and the gate network also has two layers with a hidden dimension of 128. For Cityscapes, I^{2}Net has four I^{2}Layers with hidden dimensions of 512, 256, and 256, and the gate network has three layers with hidden dimensions of 256 and 128.

3.2 Results and Analysis

Model scaling by *K*. We first explore the impact of the critical hyperparameter K on the performance of I^{2}Net. Results are shown in Table 1 in supplementary material. We observe that I^{2}Net achieves excellent performance when K reaches 3 or 4.

Comparison with aligning methods. To verify that our I^{2}Net indeed better exploits the misaligned contextual codes, we compare our method against three groups of methods, i.e. recent INR-based methods (IFA [11], IOSNet [15]), naive aligning methods (interpolation methods and DeconvNet [20]), fancy aligning methods (including state-of-the-art methods like AlignSeg [12] and SFNet [17]). As shown in Table 1, our I^{2}Net achieves the best performance over all the baseline methods. Moreover, our I^{2}Net brings much fewer overheads to the decoder than some state-of-the-art aligning methods, i.e. AlignSeg and SFNet. Thus, our I^{2}Net is a simple but effective method, which achieves a better trade-off between computational cost and accuracy than all the previous methods. We further visualize some results in Fig. 3 to show the superiority of our I^{2}Net.

Table 1. Comparison with different aligning methods on Glas test and Synapse val. The best results are in **boldface** and the second best underlined.

Method	Glas				Synapse											
	#Params	GFLOPs	DSC(%)	HD95(mm)	#Params	GFLOPs	DSC(%)	HD95(mm)	Aorta	Gallbladder	Kidney(L)	Kidney(R)	Liver	Pancreas	Spleen	Stomach
Bilinear Up-sampling	11.24M	10.42	87.98	18.31	15.31M	16.11	73.24	39.06	82.23	63.31	78.92	70.88	89.15	48.21	82.27	70.96
Nearest Neighbor	11.24M	10.42	87.02	18.65	15.31M	16.11	72.44	39.23	82.84	61.83	77.44	70.89	89.77	46.79	82.08	67.89
Deconvolution [20]	12.93M	16.33	87.50	18.15	16.29M	26.48	73.21	37.08	84.41	56.57	80.54	72.93	89.46	49.11	83.40	69.28
UNet [21]	14.33M	10.46	88.99	18.51	17.26M	30.66	76.85	39.70	89.07	69.72	77.77	68.60	93.43	53.98	86.67	75.58
CARAFE++ [27]	13.81M	24.25	91.64	11.75	17.46M	41.88	76.79	29.02	85.60	67.47	78.88	72.97	92.27	**62.53**	84.62	70.00
SFNet [17]	22.40M	51.51	92.17	9.33	19.77M	177.73	77.92	29.81	86.43	65.70	82.70	79.85	92.44	56.64	85.66	73.95
AlignSeg [12]	19.13M	68.21	92.31	7.99	21.61M	194.78	77.42	29.6	87.95	63.54	**84.13**	79.46	94.04	54.45	86.92	68.90
IFA [11]	11.79M	16	90.67	12.52	15.51M	24.32	76.23	32.54	85.84	63.23	80.21	74.00	93.11	53.03	**87.71**	72.72
IOSNet [15]	-	-	-	-	15.49M	23.08	75.56	29.17	85.17	65.49	81.00	75.35	92.79	50.59	84.03	70.02
I²Net ($K = 3$)	11.40M	12.30	**93.91**	4.18	17.43M	32.53	**79.59**	**25.99**	88.56	66.55	83.99	**81.17**	**94.24**	59.10	87.17	**75.91**
I²Net ($K = 4$)	11.44M	12.99	93.84	**3.55**	17.47M	32.99	78.99	28.70	**89.44**	**69.89**	83.98	79.90	92.98	55.78	86.71	73.27

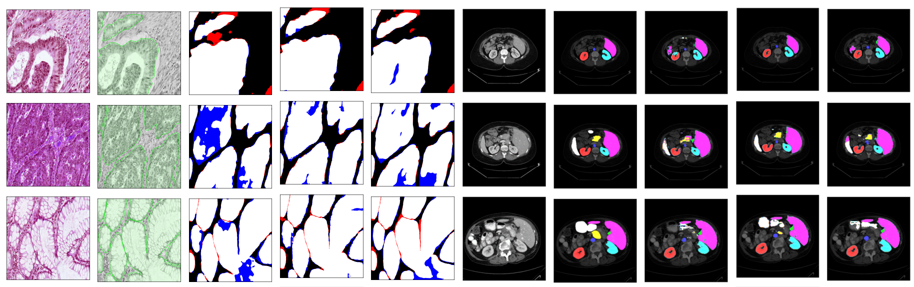


Fig. 3. Qualitative comparison on Glas test and Synapse val. 'GT' indicates groundtruth. In Glas, white pixel denotes positive, black denotes background (negative), red denotes false positive, and blue denotes false negative.

Comparison with state-of-the-arts. We further compare our I^{2}Net with the state-of-the-art methods on the Glas test and Synapse val in Table 2. In Glas, our I^{2}Net achieves the best performance with a simple ResNet-18 backbone over the advanced CNN-based methods (e.g. AttnUNet [22], PraNet [7]), the advanced generic semantic segmentation methods with well-pretrained backbones (e.g. SegFormer [28], SETR-PUP [30]), the Transformer-based methods tailored for medical data(e.g. Swin-UNet [2], MedT [24]), and the state-of-the-art methods using hybrid backbones based on Transformer and CNNs (e.g. ConTrans [18], TransFuse [29]). In Synapse, our I^{2}Net also achieves the best performance over the advanced CNN-based methods (e.g. ResUNet [6]) and the state-of-the-art methods with stronger fancy backbones (e.g. MT-UNet [26], UCTransNet [25]).

Ablation studies. To demonstrate the contribution of each component, we conduct ablation studies on I^{2}Net ($K = 3$). As shown in Table 3, introducing implicit parameterization brings a performance boost of about **2%** for DSC, indicating that it is the most critical component of our method. For gate shaping, solely incorporating $\mathcal{L}_{ges}$ or $\mathcal{L}_{gis}$ both bring performance drops to vanilla I^{2}Net, whereas utilizing them together brings a DSC improvement of about 0.5∼0.7%.

Table 2. Comparison with the state-of-art methods on Glas test and Synapse val. The best results are in **boldface** and the second best underlined.

Glas DSC(%)	Glas Backbone	Glas Method	Synapse Method	Synapse Backbone	DSC(%)	HD95(mm)	Aorta	Gallbladder	Kidney(L)	Kidney(R)	Liver	Pancreas	Spleen	Stomach
88.99	ResNet-18	[21] UNet	UNet [21]	FCN	76.85	39.70	89.07	69.72	77.77	68.60	93.43	53.98	86.67	75.58
89.98	ResNet-50	[31] UNet++	V-Net [19]	3D FCN	68.81	-	75.34	51.87	77.10	80.75	87.84	40.05	80.56	56.98
89.02	ResNet-34	[9] CENet	UNet++ [31]	ResNet-50	76.91	36.93	88.19	68.89	81.76	75.27	93.01	58.20	83.44	70.52
87.68	FCN	[22] AttnUNet	R50 UNet [3]	ResNet-50	74.68	36.87	84.18	62.84	79.19	71.29	93.35	48.23	84.41	73.92
87.49	DResNet-50	[4] DeepLabV3	AttnUNet [22]	FCN	77.77	36.02	**89.55**	68.88	77.98	71.11	93.57	58.04	87.30	75.75
91.20	Res2Net-50	[7] PraNet	R50 AttnUNet [3]	ResNet-50	75.57	36.97	55.92	63.91	79.20	72.71	93.56	49.37	87.19	74.95
89.73	MiT-B2	[28] SegFormer	DARR [8]	3D FCN	69.77	-	74.74	53.77	72.31	73.24	94.08	54.18	**89.90**	45.96
88.75	T-Base	[30] SETR-PUP	ResUNet [6]	ResUNet-a	76.95	38.44	87.06	66.05	83.43	76.83	93.99	51.86	85.25	70.13
82.52	GAT	[24] MedT	MultiResUNet [13]	MultiRes-CNN	77.42	36.84	87.73	65.67	82.08	70.43	93.49	**60.09**	85.23	75.66
88.94	Swin-B	[2] Swin-UNet	ViT [3]	ViT	61.50	39.61	44.38	39.59	67.46	62.94	89.21	43.14	75.45	69.78
90.71	ResNet-34 & MCT	[14] MCTrans	R50 ViT [3]	R50-ViT	71.29	32.87	73.73	55.13	75.80	72.20	91.51	45.99	81.99	73.95
89.94	R50-ViT	[3] TransUNet	TransUNet [3]	R50-ViT	77.48	31.69	87.23	63.13	81.87	77.02	94.08	55.86	85.08	75.62
90.18	ResNet-50 & CCT	[25] UCTransNet	TransNorm [1]	FCN & ViT	78.40	30.25	86.23	65.10	82.18	78.63	94.22	55.34	89.50	76.01
90.79	ResNet-34 & DeiT-S	[29] TransFuse	UCTransNet [25]	ResNet-50 & CCT	78.23	26.75	88.86	66.97	80.18	73.17	93.16	56.22	87.84	**79.43**
92.06	Swin-B & DAB-CNN	[18] ConTrans	MT-UNet [26]	MTM	78.59	26.59	87.92	64.99	81.47	77.29	93.06	59.46	87.75	76.81
93.91	ResNet-18	**I²Net** ($K = 3$)		FCN	**79.59**	**25.99**	88.56	66.55	**83.99**	**81.17**	**94.24**	59.10	87.17	75.91
94.00	ResNet-18	($K = 6$) **I²Net**	**I²Net** ($K = 4$)	FCN	78.99	28.70	89.44	**69.89**	83.98	79.90	92.98	55.78	86.71	73.27

Table 3. Ablation studies on Glas test and Synapse val ($K = 3$).

I²Net	Orthogonal Exploitation $\mathcal{L}_{gis}$	$\mathcal{L}_{ges}$	$\mathcal{L}_{lo}$	Glas DSC(%)	Glas HD95(mm)	Synapse DSC(%)	HD95(mm)	Aorta	Gallbladder	Kidney(L)	Kidney(R)	Liver	Pancreas	Spleen	Stomach
				90.32	12.92	76.23	32.54	85.84	63.23	80.21	74.00	93.11	53.03	**87.71**	72.72
✔				92.55	8.15	78.79	31.03	87.49	**69.48**	80.98	75.15	94.16	64.02	86.61	72.41
✔	✔			92.17	9.33	77.77	32.48	87.09	65.37	82.99	76.40	93.47	58.05	86.34	72.44
✔		✔		92.53	8.14	78.70	31.21	88.14	67.70	83.03	76.49	93.40	56.06	87.29	**77.48**
✔	✔	✔		93.22	7.06	79.26	29.81	88.05	67.11	83.71	81.04	93.73	58.24	86.97	75.24
✔			✔	93.02	6.65	79.02	28.97	87.70	67.00	83.92	80.67	93.69	58.16	85.84	75.15
✔	✔	✔	✔	**93.91**	**4.18**	**79.59**	**25.99**	**88.56**	66.55	**83.99**	**81.17**	**94.24**	**59.10**	87.17	75.91

In addition, I²Net trained without $\mathcal{L}_{gs}$ and the one without $\mathcal{L}_{lo}$ both obtain a lower DSC of about 0.4∼0.9% than the full one.

Visualization of gates. To unveil what pattern each PL learns in I²Net, we visualize some gates of I²Net ($K = 3$) trained on Synapse in Fig.1 in supplementary material.

Generalization to generic semantic segmentation. To show the generalization ability of our I²Net, we further evaluate models on a popular benchmark, i.e. Cityscapes, for generic semantic segmentation (Table 2, 3 in supplementary material). We observe the consistent superiority of I²Net over the fancy aligning methods (e.g. AlignSeg, SFNet), advanced CNN-based methods (e.g. DANet, GCNet), and state-of-the-art methods using stronger backbones (e.g. OCRNet, SETR).

4 Conclusion

In this paper, we propose I²Net, a novel implicit-parameterized INR network to capture various patterns behind contexts for medical image segmentation. We further propose novel gate shaping and learner orthogonalization to induce I²Net to learn orthogonal context patterns. Extensive experiments show that our I²Net, as a simple INR-based method, achieves superior performance over various competing methods, including fancy context aligning methods, advanced CNN-based methods, and state-of-the-art methods using stronger backbones.

Acknowledgments. This research was supported by the National Natural Science Foundation of China (Grant No.61972221).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Azad, R., Al-Antary, M.T., Heidari, M., Merhof, D.: Transnorm: Transformer provides a strong spatial normalization mechanism for a deep segmentation model. IEEE Access **10**, 108205–108215 (2022)
2. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swin-unet: Unet-like pure transformer for medical image segmentation. In: European conference on computer vision. pp. 205–218. Springer (2022)
3. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
4. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017)
5. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3213–3223 (2016)
6. Diakogiannis, F.I., Waldner, F., Caccetta, P., Wu, C.: Resunet-a: A deep learning framework for semantic segmentation of remotely sensed data. ISPRS Journal of Photogrammetry and Remote Sensing **162**, 94–114 (2020)
7. Fan, D.P., Ji, G.P., Zhou, T., Chen, G., Fu, H., Shen, J., Shao, L.: Pranet: Parallel reverse attention network for polyp segmentation. In: Medical image computing and computer-assisted intervention. pp. 263–273 (2020)
8. Fu, S., Lu, Y., Wang, Y., Zhou, Y., Shen, W., Fishman, E., Yuille, A.: Domain adaptive relational reasoning for 3d multi-organ segmentation. In: Medical Image Computing and Computer Assisted Intervention. pp. 656–666. Springer (2020)
9. Gu, Z., Cheng, J., Fu, H., Zhou, K., Hao, H., Zhao, Y., Zhang, T., Gao, S., Liu, J.: Ce-net: Context encoder network for 2d medical image segmentation. IEEE transactions on medical imaging **38**(10), 2281–2292 (2019)
10. Han, Y., Huang, G., Song, S., Yang, L., Wang, H., Wang, Y.: Dynamic neural networks: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(11), 7436–7456 (2022)
11. Hu, H., Chen, Y., Xu, J., Borse, S., Cai, H., Porikli, F., Wang, X.: Learning implicit feature alignment function for semantic segmentation. In: Proceedings of the European conference on computer vision (ECCV). vol. 13689, pp. 487–505 (2022)
12. Huang, Z., Wei, Y., Wang, X., Liu, W., Huang, T.S., Shi, H.: Alignseg: Feature-aligned segmentation networks. IEEE Trans. Pattern Anal. Mach. Intell. **44**(1), 550–557 (2022)
13. Ibtehaz, N., Rahman, M.S.: Multiresunet: Rethinking the u-net architecture for multimodal biomedical image segmentation. Neural networks **121**, 74–87 (2020)

14. Ji, Y., Zhang, R., Wang, H., Li, Z., Wu, L., Zhang, S., Luo, P.: Multi-compound transformer for accurate biomedical image segmentation. In: Medical Image Computing and Computer Assisted Intervention. pp. 326–336 (2021)
15. Khan, M.O., Fang, Y.: Implicit neural representations for medical imaging segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 433–443. Springer (2022)
16. Landman, B., Xu, Z., Igelsias, J.E., Styner, M., Langerak, T., Klein, A.: Segmentation outside the cranial vault challenge. In: MICCAI: Multi Atlas Labeling Beyond Cranial Vault-Workshop Challenge (2015)
17. Li, X., You, A., Zhu, Z., Zhao, H., Yang, M., Yang, K., Tan, S., Tong, Y.: Semantic flow for fast and accurate scene parsing. In: Proceedings of the European conference on computer vision (ECCV). vol. 12346, pp. 775–793 (2020)
18. Lin, A., Xu, J., Li, J., Lu, G.: Contrans: Improving transformer with convolutional attention for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 297–307 (2022)
19. 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)
20. Noh, H., Hong, S., Han, B.: Learning deconvolution network for semantic segmentation. In: Proceedings of the IEEE international conference on computer vision. pp. 1520–1528 (2015)
21. 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. vol. 9351, pp. 234–241 (2015)
22. Schlemper, J., Oktay, O., Schaap, M., Heinrich, M., Kainz, B., Glocker, B., Rueckert, D.: Attention gated networks: Learning to leverage salient regions in medical images. Medical image analysis **53**, 197–207 (2019)
23. Sirinukunwattana, K., Pluim, J.P., Chen, H., Qi, X., Heng, P.A., Guo, Y.B., Wang, L.Y., Matuszewski, B.J., Bruni, E., Sanchez, U., et al.: Gland segmentation in colon histology images: The glas challenge contest. Medical image analysis **35**, 489–502 (2017)
24. Valanarasu, J.M.J., Oza, P., Hacihaliloglu, I., Patel, V.M.: Medical transformer: Gated axial-attention for medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention. pp. 36–46 (2021)
25. Wang, H., Cao, P., Wang, J., Zaiane, O.R.: Uctransnet: rethinking the skip connections in u-net from a channel-wise perspective with transformer. In: Proceedings of the AAAI conference on artificial intelligence. vol. 36, pp. 2441–2449 (2022)
26. Wang, H., Xie, S., Lin, L., Iwamoto, Y., Han, X.H., Chen, Y.W., Tong, R.: Mixed transformer u-net for medical image segmentation. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 2390–2394. IEEE (2022)
27. Wang, J., Chen, K., Xu, R., Liu, Z., Loy, C.C., Lin, D.: CARAFE++: unified content-aware reassembly of features. IEEE Trans. Pattern Anal. Mach. Intell. **44**(9), 4674–4687 (2022)
28. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. Advances in Neural Information Processing Systems **34**, 12077–12090 (2021)
29. Zhang, Y., Liu, H., Hu, Q.: Transfuse: Fusing transformers and cnns for medical image segmentation. In: International Conference on Medical Image Computing and Computer Assisted Intervention. pp. 14–24 (2021)

30. Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P.H., et al.: Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6881–6890 (2021)
31. Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J.: Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. IEEE transactions on medical imaging **39**(6), 1856–1867 (2019)