ClassFormer: Exploring Class-Aware Dependency with Transformer for Medical Image Segmentation

Huimin Huang[1], Shiao Xie[1†], Lanfen Lin[1*], Ruofeng Tong[1,2], Yen-Wei Chen[3], Hong Wang[4], Yuexiang Li[4], Yawen Huang[4*], Yefeng Zheng[4]

[1] Zhejiang University
[2] Zhejiang Lab
[3] Ritsumeikan University
[4] Tencent Jarvis Lab

Abstract

Vision Transformers have recently shown impressive performances on medical image segmentation. Despite their strong capability of modeling long-range dependencies, the current methods still give rise to two main concerns in a class-level perspective: (1) intra-class problem: the existing methods lacked in extracting class-specific correspondences of different pixels, which may lead to poor object coverage and/or boundary prediction; (2) inter-class problem: the existing methods failed to model explicit category-dependencies among various objects, which may result in inaccurate localization. In light of these two issues, we propose a novel transformer, called ClassFormer, powered by two appealing transformers, i.e., intra-class dynamic transformer and inter-class interactive transformer, to address the challenge of fully exploration on compactness and discrepancy. Technically, the intra-class dynamic transformer is first designed to decouple representations of different categories with an adaptive selection mechanism for compact learning, which optimally highlights the informative features to reflect the salient keys/values from multiple scales. We further introduce the inter-class interactive transformer to capture the category dependency among different objects, and model class tokens as the representative class centers to guide a global semantic reasoning. As a consequence, the feature consistency is ensured with the expense of intra-class penalization, while inter-class constraint strengthens the feature discriminability between different categories. Extensive empirical evidence shows that ClassFormer can be easily plugged into any architecture, and yields improvements over the state-of-the-art methods in three public benchmarks.

1 Introduction

Medical image segmentation is crucial to many clinical applications, such as diagnosing disease and preoperative assessment. Over the last decade, Convolutional Neural Networks (CNNs) have greatly promoted the development of medical image segmentation (Ronneberger, Fischer, and Brox 2015; Zhou et al. 2019). Despite their success, CNNs learn a limited receptive field from the convolution filters, which fail to model the explicit long-range relations (in Fig. 2 (d)). Rather, the introduction of large, sometimes even global contexts endows models with richer representations, especially for medical images with complex anatomical contrasts, anfractuous boundaries, and heterogeneous textures.

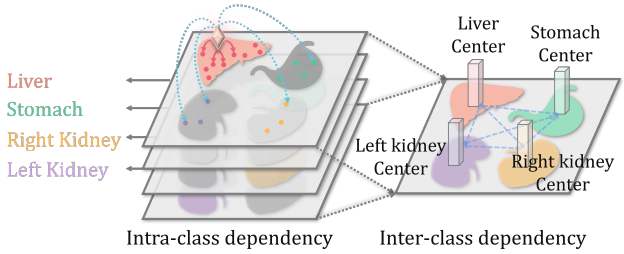

Figure 1: Motivation of ClassFormer. Intra-class dependency considers the highly-structured patterns for each object, *e.g.,* liver region is relatively highlighted in its class-specific representation, and then the pixel-to-pixel relations are explored. Inter-class dependency further considers the high-level object-to-object relations (*e.g.,* liver-to-stomach), with additional guidance of multiple class tokens (centers).

Recently, transformers have been emerged as alternative architecture in vision tasks attracting more interest. Unlike CNN-based methods that rely on local convolution operators, visual transformers become highly considerate to learn the expressively attentive relations of different patch tokens, which are proven to be an efficient way by modeling the global contexts. For example, on the medical image segmentation task, TransUNet (Chen et al. 2021) was the first study to build a CNN-Transformer architecture, which added transformers into the high-level CNN features to learn the global attention. SwinUNet (Cao et al. 2021) was designed as a pure transformer architecture by stacking Swin transformer blocks. The global relations release the advantages of transformer-based methods, nevertheless, simply enlarging global attention still gives rise to the limited coverage of objects and inaccurate localization, as depicted in Figs. 2 (b) and (c). Substantially, the difficulty of inconsistency in the same category and the confused semantics among categories, all lead to inseparable representations.

The observed phenomenon can be elaborated as the fol-

*Corresponding Authors: Lanfen Lin (llf@zju.edu.cn), Yawen Huang (yawenhuang@tencent.com).
†Huimin Huang and Shiao Xie are co-first authors, and this work is done during the internship at Tencent Jarvis Lab.

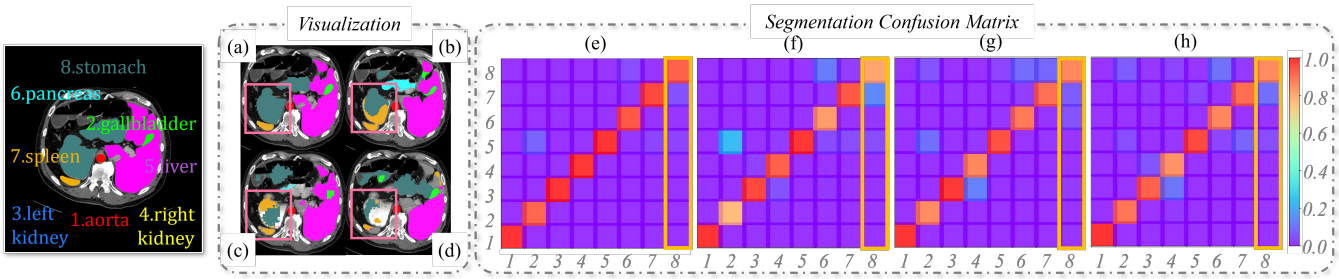

Figure 2: Visualization of eight-organ segmentation and confusion matrix with column normalization, (a)(e) our ClassFormer, (b)(f) TransUNet, (c)(g) SwinUNet, (d)(h) ResUNet. Normally, in confusion matrix, the diagonal should be brighter (intra-class compactness), while the rest should be darker (inter-class mis-segmentation). TransUNet, SwinUNet and ResUNet suffer from inter-class confusion (*e.g.,* mis-classify stomach as spleen) and intra-class incompactness (*e.g.,* incomplete stomach) in pink boxes, which can be observed at yellow boxes in (f)-(h) with darker of stomach-to-stomach and brighter of stomach-to-spleen. ClassFormer improves the transformer from a class-level perspective, achieving accurate localization and better compactness.

lowing two class-level perspectives. (**i**) ***Intra-class dependency***: The existing transformer-based methods typically reason on a class-agnostic feature map, where all foreground objects are highlighted simultaneously. The weighted aggregation of all such salient regions (reasoned on multiple objects) leads to the dubious pixel-to-pixel relations, which may harm the consistent learning of the intra-class representations, *e.g.*, the incomplete stomach segmentation shown in Fig. 2. (**ii**) ***Inter-class dependency***: Recent transformers usually focus on pixel-to-pixel dependencies, which may not be an ideal design, considering the most-overlooked aspect of object-to-object correlation between various semantic categories. Neglecting inter-class dependencies limits the ability for accurate segmentation between different organs, especially for organs with similar contextual information or/and surrounding position, *e.g.*, a part of stomach area is incorrectly segmented as spleen in Fig. 2. To address these limitations, we design a class-aware transformer for medial image segmentation, namely ClassFormer, which is motivated by above two aspects as depicted in Fig. 1.

Firstly, considering that context modeling of different objects is distinctive, especially for medical images with large organ variances (*e.g.*, shape, size), an ***intra-class dynamic transformer*** is designed to explore the class-specific pixel-to-pixel relations, which yields more consistent representation for each class. Specifically, we first decouple the representations of different organs, such that each representation can highlight the object-related regions, and relatively suppress unnecessary ones. Then, for each class, we update them separately by applying transformer with the ability of global reasoning. However, previous transformers utilize a massive number of keys to attend per query, which yields redundant computation with high complexity. To achieve this, we design a lightweight dynamic transformer, equipped with a powerful pyramid mechanism, to constrain less-relevant regions and recognize salient keys from multiple scales.

Secondly, as many studies described, the cross-category relations explicitly establish the semantic correspondence, which are capable for generating separable representations. Inspired by this, an ***inter-class interactive transformer*** is designed to capture the class-guided object-to-object correlations, via performing semantic reasoning among different object categories. Specifically, we introduce the class tokens represented by the class centers, and feed them into a transformation module together with the patch tokens with high-level semantics for interaction. In this way, class tokens are learned to guide patch tokens to model the object-to-object relations and enhance the feature discriminability.

The ability of ClassFormer in modeling class-aware dependency is shown in Fig. 2 (e). Compared with other methods, ClassFormer improves the feature correlations of the same category (brighter diagonal), as well as the feature discriminability between different classes (darker non-diagonal), achieving accurate segmentation shown in Fig. 2 (a). To summarize, our contributions are four-fold. (**1**) We analyze the intra-class and inter-class dependencies neglected by the existing transformer-based methods, and propose a plug-and-play module in a class-level perspective, called ClassFormer, to improve the performance of medical image segmentation. (**2**) We introduce an intra-class dynamic transformer to learn the pixel-to-pixel dependencies of class-specific representations. (**3**) We design an inter-class interactive transformer to model the high-level object-to-object correlations among different semantic categories. (**4**) ClassFormer is extensively tested on three datasets, which achieves state-of-the-art performances consistently, by exploring rich class-aware dependencies over local features.

2 Related Work

Vision transformers. ViT (Dosovitskiy et al. 2020) is the first study to bring Transformer into visual tasks and makes it competitive with CNNs. To reduce the quadratic computational cost of self-attention, Swin Transformer restricts attention in local windows. PVT (Wang et al. 2021) directly downsamples the number of keys and values, while DAT (Xia et al. 2022) focuses on important regions by learning offset of each reference point. Recently, a series of transformer-based backbones, e.g. TransUNet and SwinUNet, have been proposed for medical image segmentation and gained strong performance. However, they neglected the

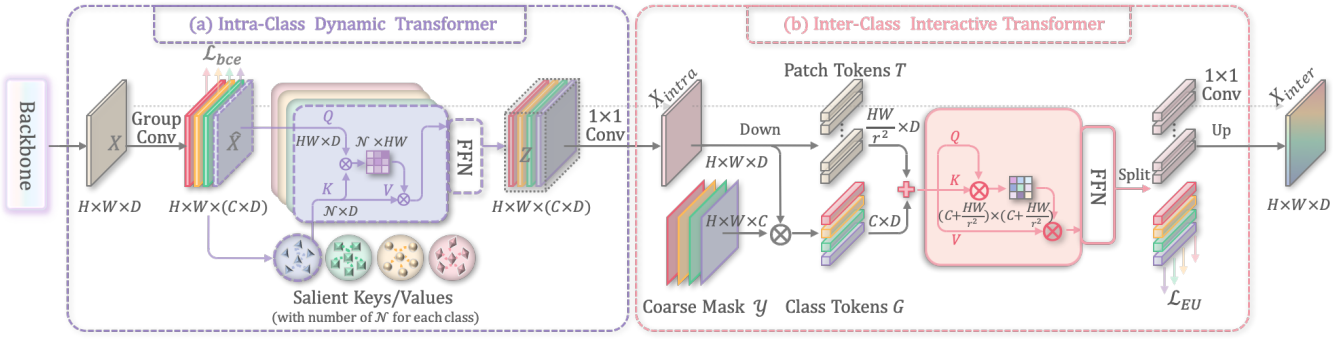

Figure 3: An overview of ClassFormer. It has two sequential sub-modules: (a) intra-class dynamic transformer and (b) inter-class interactive transformer, which can be plugged into any stage (in encoder or decoder) of backbone to improve its performance.

class-aware dependency. In contrast, our ClassFormer improves the transformer in the class-level perspective, which improves the integrity of target segmentation and reduces class confusion with a relatively small computational cost.

Class-wise reasoning. Recently, the class-wise dependency has been proven to play a critical role in semantic segmentation. Both of OCR (Yuan, Chen, and Wang 2020) and ACFNet (Zhang et al. 2019) aggregate the class centers by weighted average of the coarse mask and pixel features. Different from ACFNet obtaining the contextual feature by using probability map, OCR calculates the similarity between pixel features and class centers. Inspired by OCR, MCIBI (Jin et al. 2021) further builds a feature memory module to store the dataset-level features of various categories. CDGCNet (Hu et al. 2020) introduces the dynamic graph convolution that allocates attention to pixels in the same category. However, these methods simply consider either intra-class or inter-class dependency. In our work, we establish both two dependencies to enhance features for better performance.

3 Method

3.1 Overview of ClassFormer

As shown in Fig. 3, the input image is firstly fed to the backbone network to generate the class-agnostic representation X. Then, an intra-class dynamic transformer (in Sec. 3.2) is implemented to explore the pixel-to-pixel relations on each class-specific representation, instead of directly reasoning on the class-agnostic feature map. To further learn the correlation between different semantic categories, we propose the inter-class interactive transformer (in Sec. 3.3), which exploits the relations between patch tokens and the newly introduced class tokens (class centers). Finally, the separable representations from two successive transformers are sent to the downstream blocks for pixel-wise prediction. An in-depth discussion on ClassFormer is conducted along with the analysis of its complexity (in Sec. 3.4). In the following sections, we will introduce our ClassFormer in details.

3.2 Intra-class Dynamic Transformer

Class-specific representation with supervision. Given an input image I, we first utilize a backbone network (*e.g.,* ReNet) to extract the feature map $X \in \mathbb{R}^{H\times W\times D}$, where H, W, and D are the corresponding height, width, and feature dimension. After that, the class-agnostic feature map is repeated for C times (*i.e.* $X \in \mathbb{R}^{H\times W\times(C\times D)}$), where C is the number of categories, and then the repeated feature map is fed to a group convolutional layer with C groups to generate a collection of decoupled feature maps $\hat{X} = [X_1, X_2, \cdots, X_c]$, where $X_c \in \mathbb{R}^{H\times W\times D}$ is the feature for the c-th class and $[\cdot]$ denotes the concatenation operation.

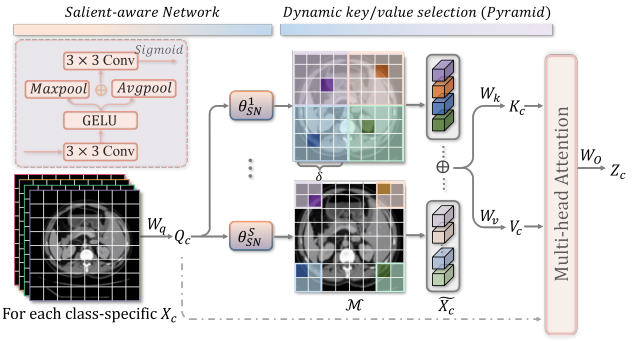

Figure 4: An illustration of intra-class dynamic transformer.

To explicitly integrate the class-aware cues into the decoupled features, each feature map is predicted as a binary mask, *i.e.* the pixel value means the possibilities of the pixel belonging to the category. Hence, a binary cross-entropy loss $\mathcal{L}_{bce}$ can be calculated using the one-hot encoded ground truth, which gradually makes the features class-awareness.

Dynamic Transformer. To capture the class-specific long-range dependency, a straightforward method is to apply transformer with global relations for each class. However, in the conventional transformer, each query patch attends to a large number of keys, resulting in the redundant calculations of correlations among irrelevant keys. Hence, a dynamic transformer is designed with saliency-aware keys/values for the given class-specific query, which can adaptively learn the visual pattern for individual class with an efficient trade-off. Note that the salient keys/values are selected from not only the target category (*i.e.* salient positives), but also other cate-

gories (*i.e.* salient negatives). In this way, the salient positive samples can guide the learning of the salient negatives, leading to to a better extraction of feature representations.

As illustrated in the Fig. 4, given the input feature map $X_c \in \mathbb{R}^{H\times W\times D}$, the query tokens can be obtained via a linear projection: $Q_c = X_c W_q$, which are subsequently fed to a simple ***Saliency-aware Network***, θ_{SN} (left corner in Fig. 4) to re-weight the pixels, *i.e.* increasing the weights for salient pixels while suppressing the irrelevant ones. In the saliency-aware network, input features first go through a 3×3 convolution operation and GELU activation for the extraction of local features. Since pooling along the channel dimension can effectively activate the informative regions (Woo et al. 2018; Huang et al. 2021a), we perform the channel-wise max- and average-pooling, and then concatenate them. The concatenated feature is then processed by a 3×3 convolutional layer to yield the saliency-aware map $\mathcal{M} = \theta_{SN}(Q_c) \in \mathbb{R}^{H\times W}$.

To achieve ***Dynamic Key/Value Selection*** in the whole space Λ, we first divide the saliency-aware map $\mathcal{M}$ into a set of non-overlapping sub-regions Λ_s, by using the sliding window with size of δ. Within each $\Lambda_s \in \mathbb{R}^{\delta\times\delta}$, the salient position with maximum localization probability is selected:

$$\hat{p} = \arg\max_{p\in\Lambda_s} \mathcal{M}(p), \tag{1}$$

where p is the position of each token. Accordingly, a set of salient positions $P = \{\hat{p}_n\}_{n=1}^{N}$ is formed for the whole space Λ, where N represents the number of positions and equals to $\lceil H/\delta \rceil \times \lceil W/\delta \rceil$ ($\lceil\cdot\rceil$ is the ceiling operation).

Since the finer sampling with a smaller receptive scale (smaller δ) is observed to focus on the tiny objects with explicit long-range context, while the coarser sampling with a larger receptive field (larger δ) concentrates on the large objects with global dependencies, a ***Pyramid Mechanism*** is implemented to adaptively sample salient positions from multiple scales. Concretely, for each scale, we utilize different saliency-aware networks θ_{SN} with corresponding δ (the size of sliding window), which can adaptively learn the distinctive patterns for varying scales. Thereby, the pyramid $\tilde{P}$ fuses the information extracted from varied scales: $\tilde{P} = \{P_s\}_{s=1}^{S}$, where S is total number of scales.

According to the salient positions $\tilde{P}$, we can sample the salient tokens $\tilde{X}_c$ from the input X_c. The keys K_c and values V_c can be obtained by linear projection:

$$K_c = \tilde{X}_c W_k, V_c = \tilde{X}_c W_v. \tag{2}$$

With the linearly projected Q_c, K_c and V_c, we perform the multi-head attention to capture long-range dependency:

$$Z_c = softmax(Q_c {K_c}^T/\sqrt{D})V_c. \tag{3}$$

The enhanced sequence $Z = \{Z_c\}_{c=1}^{C}$ from different classes is reshaped into a $2D$ feature and concatenated with the input feature, which is then sent to a 1×1 convolutional layer for the generation of the final output $X_{intra} \in \mathbb{R}^{H\times W\times D}$.

3.3 Inter-class Interactive Transformer

Category-wise dependency has been proven to be essential for context modeling, which aims at exploring the high-level semantic relationship among local regions. Previous transformers perform reasoning on the feature maps with fine-grained details (in Fig. 5), which only focuses on the pixel-level relations. Such a setting limits the potential of modeling explicit category dependency among semantic objects. To tackle this issue, we explore object-to-object relations via the interactions among patch tokens with high-level semantics and multiple class-distinctive tokens. In the next section, we will describe the semantic attention process in details.

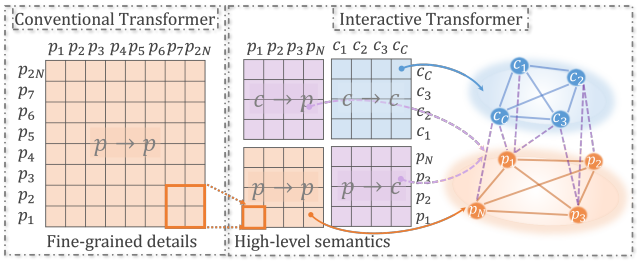

Figure 5: Interactions among patch (p_i) and class tokens (c_i).

Patch tokens (T)**.** Given the input $X_{intra} \in \mathbb{R}^{H\times W\times D}$, we first utilize a max-pooling with stride r and two successive 3×3 convolutional layers to yield a high-level representation of the size $H/r \times W/r \times D$. Then, we further reshape it into a sequence of patch tokens $T \in \mathbb{R}^{HW/r^2\times D}$ with rich semantics. In the experiment, the model achieves the best performance when the number of patch tokens HW/r^2 is close to the class number C (in Sec. 4.2).

Class tokens (G)**.** A set of new tokens, namely class tokens (represented by class centers), is integrated into transformer. The center of each class can be calculated by taking the average of all feature pixels of the same class. Hence, we first utilize X_{intra} to predict a coarse segmentation probability map $\mathcal{Y}_c \in \mathbb{R}^{H\times W}$ ranging from 0 to 1 for each class. Then, we aggregate the representations of all pixels weighted by their probabilities of belonging to the c-th object:

$$G_c = \sum_{i\in I} \mathcal{Y}_c^i X_{intra}^i, \tag{4}$$

where X_{intra}^i is the feature of i-th pixel and $\mathcal{Y}_c^i$ is the normalized probability for i-th pixel belonging to the c-th class. We denote the obtained $G_c \in \mathbb{R}^D$ as the c-th class token.

Interactive Transformer. As shown in Fig. 5, we concatenate class tokens $G \in \mathbb{R}^{C\times D}$ with patch tokens $T \in \mathbb{R}^{HW/r^2\times D}$, which then passes through the transformer to capture the long-range interactions. The linearly projected Q, K and V have the same shape of $(C + HW/r^2) \times D$, and the attention map $\mathcal{A}$ can be denoted as:

$$\mathcal{A} = \begin{bmatrix} \mathcal{A}^{c\to c} & \mathcal{A}^{c\to p} \\ \mathcal{A}^{p\to c} & \mathcal{A}^{p\to p} \end{bmatrix}, \tag{5}$$

where $\mathcal{A}^{c\to c}$, $\mathcal{A}^{p\to p}$, $\mathcal{A}^{c\to p}$, $\mathcal{A}^{p\to c}$ are normalized similarity matrices representing class-to-class, patch-to-patch, class-to-patch, and patch-to-class. $\mathcal{A}^{c\to c}$ and $\mathcal{A}^{p\to p}$ perform the

Method	Params	DSC	HD	Aorta	Gallbladder	Kidney(L)	Kidney(R)	Liver	Pancreas	Spleen	Stomach
V-Net	45.60	68.81	-	75.34	51.87	77.10	80.75	87.84	40.05	80.56	56.98
U-Net	31.04	76.85	39.70	89.07	69.72	77.77	68.60	93.43	53.98	86.67	75.58
AttUNet	34.88	77.77	36.02	89.55	68.88	77.98	71.11	93.57	58.04	87.30	75.75
R50 ViT	103.77	71.29	32.87	73.73	55.13	75.80	72.20	91.51	45.99	81.99	73.95
TransUNet	105.28	77.48	31.69	87.23	63.13	81.87	77.02	94.08	55.86	85.08	75.62
SwinUNet	27.17	79.12	21.55	85.47	66.53	83.28	79.61	94.29	56.58	90.66	76.60
UCTransNet	66.43	78.23	26.75	88.86	66.97	80.19	73.18	93.17	56.22	87.84	79.43
MTUnet	79.08	78.59	26.59	87.92	64.99	81.47	77.29	93.06	59.46	87.75	76.81
AFTerUNet	41.50	81.02	-	**90.91**	64.81	87.90	85.30	92.20	63.54	90.99	72.48
MISSFormer	42.46	81.96	18.20	86.99	68.65	85.21	82.00	94.41	65.67	**91.92**	80.81
ScaleFormer	111.58	82.86	16.81	88.73	**74.97**	86.36	83.31	95.12	64.85	89.40	80.14
ClassFormer	35.43	**84.40**	**12.22**	89.04	72.91	**88.95**	**85.68**	**95.55**	**69.25**	90.70	**83.09**

Table 1: Comparison with SOTA methods on Synapse. ClassFormer is plugged into the deepest stage of the ResUNet decoder.

self-attentions, which update the class tokens and patch tokens respectively; $\mathcal{A}^{c\to p}$ and $\mathcal{A}^{p\to c}$ learn the complementary cues via cross-attention. Hence, patch tokens can explore the explicit category dependencies from class tokens, resulting in the class-guided feature enhancement; while class tokens can capture rich semantics from patch tokens, which can improve the discriminability of various categories.

After that, we reverse the enhanced sequence back to class tokens and patch tokens according to the order of concatenation. Moreover, we introduce an auxiliary Euclidean distance loss $\mathcal{L}_{EU}$ (Li et al. 2022) to maximize the distance between the enhanced class tokens, which can further improve the feature discrepancy. The enhanced patch tokens are reshaped into $2D$ features and upsampled into the original size of $H \times W \times D$. To retain the source information, the input X_{intra} is also combined, which is reformed by a 1×1 convolutional layer, to obtain the final output $X_{inter} \in \mathbb{R}^{H\times W\times D}$.

3.4 ClassFormer

Arrangement of transformer modules. Given an input feature map, two transformer modules, *i.e.* intra-class and inter-class, calculate the complementary dependencies. Actually, we can place these two transformers in a sequential or parallel manner (in Sec. 4.5) for feature extraction. The experimental results reveal that the sequential arrangement outperforms the parallel arrangement, where the intra-first order is better than the inter-first. In the intra-first order, the enhanced features can first focus on the internal object structure, and then provide the inter-class transformer with more compact feature for each object and more representative initial class tokens, which achieves the goal of comprehensively exploring class-aware dependency.

Complexity of ClassFormer. Compared with the conventional transformer with the complexity of $O(DH^2W^2)$, the cost of ClassFormer is decreased and can be summarized as:

$$\Omega = \underbrace{O(DHWC\mathcal{N})}_{intra-class} + \underbrace{O(D(C+HW/r^2)^2)}_{inter-class}. \quad (6)$$

In the intra-class dynamic transformer, $\mathcal{N} = \sum_{i=1}^{S} N_i$ is the total number of selected salient keys from S scales. Even though multiple classes C and scales S are taken into account, the cost of intra-class transformer is comparably small to the conventional one ($C\mathcal{N} \ll HW$). Additionally, in the inter-class interactive transformer, the down-sampling rate r determines the number of high-level patch tokens, which largely reduces the computational cost of transformer in a global perspective. In other words, both of our intra- and inter-class transformers are designed in a lightweight way.

4 Experiments

4.1 Datasets and Performance Metrics

Three public datasets with different numbers of categories are adopted, *i.e.*, Multi-organ Synapse dataset (Synapse), the Automated Cardiac Diagnosis Challenge dataset (ACDC), and the Multi-Organ Nucleus dataset (MoNuSeg).

Synapse (9 classes): It consists of 30 Computed Tomography (CT) scans, with 18 cases used for training and 12 cases for testing. Following (Chen et al. 2021), the Dice Coefficient (DSC) and Hausdorff Distance (HD) are averagely computed over 8 organs for quantitative comparisons.

ACDC (4 classes): It contains 100 Magnetic Resonance Imaging (MRI) scans with three organs, left ventricle (LV), right ventricle (RV), and myocardium (MYO). Following (Chen et al. 2021), we report the average DSC with a random split of 70 cases for training, 10 cases for validation, and 20 cases for testing. Since a standard division is unavailable, we could only randomly split the dataset by ourselves.

MoNuSeg (2 classes): It contains 30 images for training, and 14 images for testing. Following (Kumar et al. 2017), we report DSC and Intersection over Union (IoU).

4.2 Implementation Details

To avoid overfitting, two kinds of data augmentations, including random rotation and flipping , are adopted. Here, we list the batch size (bs), learning rate (lr), maximum training epochs (ep), optimizer (opt) for three datasets as below:

- Synapse: bs=8; lr=3e-3; ep=600; opt=SGD;
- ACDC: bs=8; lr=3e-3; ep=200; opt=SGD;
- MoNuSeg: bs=4; lr=1e-3; ep=200; opt=Adam.

All models are trained with momentum 0.9 and weight decay 1×10^{-4}. For fair comparison, most of experiments related to ClassFormer (without pre-training) are conducted on

Methods	DSC	RV	Myo	LV
R50 UNet	87.55	87.10	80.63	94.92
R50 AttUNet	86.75	87.58	79.20	93.47
R50 ViT	87.57	86.07	81.88	94.75
TransUNet	89.71	88.86	84.53	95.73
SwinUNet	90.00	88.55	85.62	95.83
TransUNet*	89.63	86.70	86.96	95.33
SwinUNet*	89.16	87.04	86.22	94.24
MTUnet*	89.84	86.06	88.14	95.32
UCTransNet*	90.11	88.01	87.29	95.02
ScaleFormer*	90.17	87.33	88.16	95.04
ClassFormer	**90.99**	**88.35**	**89.08**	**95.54**

Table 2: Comparison with SOTA methods on ACDC.

Methods	DSC	IoU
U-Net	73.97	59.42
UNet++	75.28	60.89
AttUNet	76.20	62.64
MedT	79.24	65.73
MISSFormer	76.04	61.58
MTUNet	77.97	64.30
TransUNet	79.20	65.68
SwinUNet	78.49	64.72
UCTransNet	79.87	66.68
ScaleFormer	80.06	66.87
ClassFormer	**80.75**	**67.88**

Table 3: Comparison with SOTA methods on MoNuSeg.

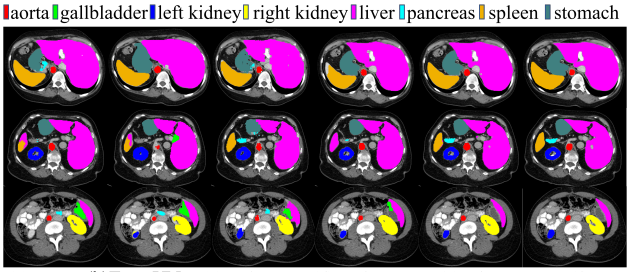

Figure 6: Visual Comparison of five methods on Synapse.

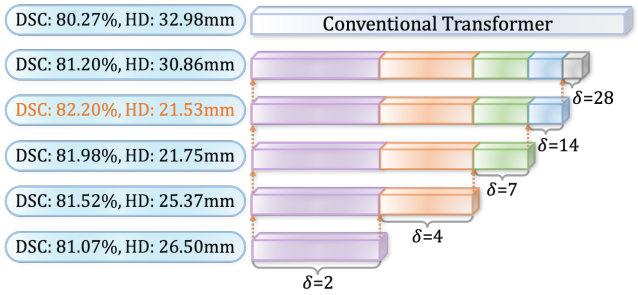

Figure 7: Applying pyramid tokens in intra-class module.

the ResUNet backbone, whose encoder has four CNN stages comprised of basic ResNet-34 blocks. In this setting, ClassFormer is integrated into the deepest stage of the ResUNet decoder with size of 28×28. Additional experiments with different backbones and stages (in Sec. 4.5) are conducted to examine the effectiveness of ClassFormer. Following (Chen et al. 2021), we use the combined cross entropy loss and dice loss as the refined and coarse segmentation losses. Additionally, $\mathcal{L}_{bce}$ and $\mathcal{L}_{EU}$ are balanced by two coefficients, λ_{bce} and λ_{EU}, which are adjusted in Sec. 4.4.

4.3 Comparison with the State-of-the-Arts

Quantitative Comparison: We first compare ClassFormer with 11 state-of-the-art (SOTA) methods on Synapse, including V-Net (Milletari, Navab, and Ahmadi 2016), U-Net, AttUNet (Schlemper et al. 2019), ViT, TransUNet, SwinUNet, UCTransNet (Wang et al. 2022a), MTUNet (Wang et al. 2022b), AFTerUNet (Yan et al. 2022), MISSFormer (Huang et al. 2021b) and ScaleFormer (Huang et al. 2022).

Table 1 lists the performance of different comparison methods on the Synapse dataset. As seen, our ClassFormer outperforms the existing approaches in both regional measures DSC (84.40%) and boundary-aware measure HD (12.22mm). In particular, ClassFormer surpasses the previous best method ScaleFormer by a large margin (DSC: +1.54% and HD: -4.59mm) with fewer model parameters (Params: -76.15M), establishing a new SOTA. Table 2 reports the DSC results of all the competing approaches on our divided ACDC dataset (*). For fair comparison, we utilize the released codes and original settings in the experiment. It is easily observed that our method achieves the best segmentation performance for the comparison methods. Table 3 presents the DSC and IoU results on the MoNuSeg dataset. Attributed to the powerful feature extraction capability, the proposed ClassFormer obtains the highest DSC and IoU scores. These quantitative results on three datasets substantiates the fine robustness of our ClassFormer.

Additionally, we conduct the statistical tests on Table. 1, 2 and 3, and the results of t-test (with $p - value$ <0.05) show that the performance benefits of our method against others are statistically significant in all cases.

Visual Comparison: Fig. 6 illustrates the qualitative results of different methods on the Synapse dataset, including U-Net, TransUNet, SwinUNet, and ScaleFormer. It demonstrates that our ClassFormer is able to precisely segment salient organs with widely variable sizes, shapes, and locations. Moreover, the object boundaries achieved by our method are clearer and sharper than other baselines.

4.4 Hyper-Parameters

Multiple Scales in Intra-class Transformer. We examine the significance of selecting salient keys among multiple scales. Each scale is determined by the size of sliding window δ (in Sec. 3.2). As seen in Fig. 7, the performance is improved with stacking scales, and the best result is achieved when using δ=2,4,7,14. However, applying the transformer on δ=28 that only selects a single key may affect the genuine characteristics. We also compare with the conventional transformer that attends on all keys with large computational

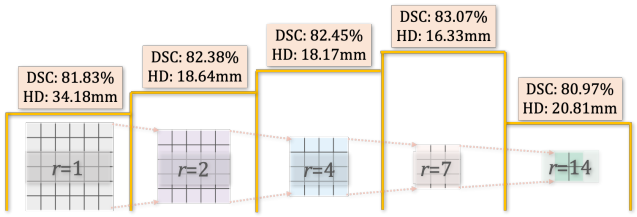

Figure 8: Global perspective in inter-class transformer.

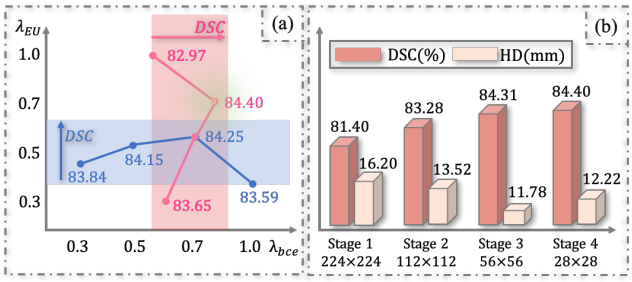

Figure 9: (a) Performance of ClassFormer w.r.t. λ_{bce} and λ_{EU}. (b) Impact of the insertion location of ClassFormer at different stages of the ResUNet encoder.

cost. We observe that dynamic transformer achieves better performance than the conventional models due to the flexibility in modeling long-range dependencies.

Global Perspective in Inter-class Transformer. The downsampling rate r (in Sec. 3.3) affects the global perspective and computational cost of the inter-class interactive transformer. The comparison results are reported in Fig. 8, where r=1 means using a fine-grained feature without downsampling. As seen, a large number of patch tokens may incur irrelevant calculation and inevitably increase the computational complexity; while extremely few number of patch tokens may result in insufficient and noisy representation. The setting r=7 achieves a good trade-off between accuracy and complexity. In this setting, the number of patch tokens is 16, which is close to the category number (9 classes) in Synapse. It indicates that the importance of modeling the category dependency in a proper global perspective.

Adjustment of λ_{bce} and λ_{EU}. Here, λ_{bce} and λ_{EU} are two coefficients that balance the overall losses. As seen in Fig. 9 (a), we first evaluate the effect of λ_{bce} with λ_{EU}=0.5 (in blue). With the increasement of λ_{bce}, the accuracy is improved. However, the accuracy starts to decline at λ_{bce}=0.7. It indicates that λ_{bce} can help the model to focus on the target objects; however, large λ_{bce} may hinder the global reasoning since the background is totally ignored. Further, we set λ_{bce}=0.7 and use different value of λ_{EU} (in gray). Overall, λ_{EU}=0.7 achieves the best results of 84.40% DSC.

4.5 Ablation Study

Effectiveness of Intra- and Inter-class Transformers. We begin by assessing the influence of intra- and inter-class transformers when integrating them into ResUNet. As seen in Table. 4, a significant improvement (DSC: +3.17%, HD: -12.33mm) is achieved when applying the intra-class dynamic transformer (#2) to the baseline ResUNet (#1). $\mathcal{L}_{bce}$ (#3) contributes to generate class-aware features and achieve better performance. Regarding to the inter-class interactive transformer, the class token (#5) introduces the guidance for inter-class reasoning, which yields an increasement of +2.31% in DSC compared to the conventional one (#4) that only performs reasoning on patch tokens. Combined with $\mathcal{L}_{EU}$ (#6) that maximizes the distance of centers, the final performance of inter-scale interactive transformer is 83.07% in DSC, which even outperforms the previous SOTA methods in Table. 1. We can also observe from the comparison results that by arranging two modules in sequential (#8, #9) is better than doing in parallel (#7). In addition, the intra-first order (#9) performed better than the inter-first order (#8), which achieves the best performance of 84.40% in DSC.

#	Intra-class		Inter-class			DSC	HD
	Trans	$\mathcal{L}_{bce}$	Trans	center	$\mathcal{L}_{EU}$	(%)	(mm)
1	×	×	×	×	×	78.69	42.32
2	✓	×	×	×	×	81.86	29.99
3	✓	✓	×	×	×	82.20	21.53
4	×	×	✓	×	×	80.46	15.38
5	×	×	✓	✓	×	82.77	15.35
6	×	×	✓	✓	✓	83.07	16.33
7	Parallel					83.70	14.53
8	Sequential: Inter→Intra					83.74	13.19
9	Sequential: Intra→Inter					**84.40**	**12.22**

Table 4: Ablation study of ClassFormer.

Backbone	ClassFormer	DSC (%)	HD (mm)
UNet	×	76.85	39.70
	✓	83.89	12.11
TransUNet	×	77.48	31.69
	✓	81.25	26.27
SwinUNet	×	79.12	21.55
	✓	80.59	27.67

Table 5: Performance of ClassFormer on various backbones.

Methods	DSC (%)	HD (mm)
baseline(ResUNet)	78.69	42.32
+ OCR	81.23	27.78
+ MCIBI	79.51	39.65
+ ACFNet	79.94	40.68
+ CDGCNet	79.45	47.37
+ ClassFormer	**84.40**	**12.22**

Table 6: Comparison with class-wise methods on Synapse.

Effectiveness of ClassFormer. To verify the capability of ClassFormer in class-wise reasoning, we compare it with OCR, MCIBI, ACFNet and CDGCNet based on the ResUNet backbone. As seen in Table. 6, ClassFormer achieves a better accuracy by comprehensively modeling class-aware

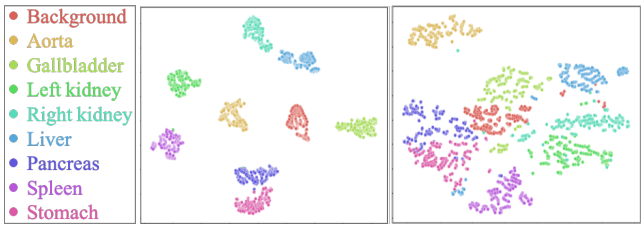

Figure 10: Visualization of deep features from ClassFormer (left) and baseline ResUNet (right) on Synapse.

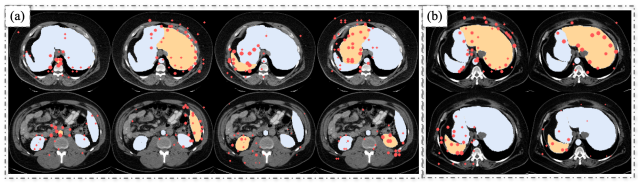

Figure 11: Most important keys selected for each organ on Synapse. The red circles with larger radii show the key points with higher scores. (a) Keys selected at scale δ=2. (b) Comparison of δ=2 (1st column) and δ=4 (2nd column).

dependency in both intra- and inter-class manners.

Different Stages. We next explore the influence of ClassFormer at different stages by integrating it into various decoder stages of ResUNet (stages 1-4), one stage at a time. As shown in Fig. 9 (b), ClassFormer brings performance benefits when introduced at each of these stages of the architecture. Especially, the best results are achieved when applied at stage 4 in the deepest layer with size of 28×28.

Different Backbones. Finally, we study the effect of integrating ClassFormer with three further representative architectures for medical image segmentation, including CNN-based U-Net, transformer-based TransUNet and SwinUNet. As reported in Table. 5, we find that ClassFormer achieves a consistent improvement in Dice and often HD across these classic architectures.

4.6 Interpretation of ClassFormer

Distribution of Deeply Learned Features. The t-SNE is used to obtain the 2D embeddings and visualize the deep features from the last encoder layer. As shown in Fig. 10, the learned pixel embeddings by ClassFormer become more compact and well separated, which indicates that the designed transformers from class-level perspective benefit the discriminative power of deeply learned features, which is crucial for medical image segmentation.

Visualization of Most Important Keys. As seen in Fig. 11, intra-class dynamic transformer learns to select salient keys from multiple scales, where most keys embed or surround the target organ. Interestingly, in Fig. 11 (b), the smaller receptive scale (δ=2 in 1st column) has a dense sampling that helps the model focus on smaller organs and the boundary of larger organs; whereas the larger receptive scale (δ=4 in 2nd column) has a sparse sampling that contributes to filtering out irrelevant keys and emphasizes on larger salient objects.

5 Conclusion

In this paper, we proposed a novel transformer-based module, termed as ClassFormer, which was able to improve the representational power of a network by enabling it to capture the class-aware dependency. Specifically, we designed two appealing transformers that simultaneously focused on both pixel-to-pixel and object-to-object dependencies. Our method was extensively evaluated on three public datasets and consistently outperformed other approaches.

6 Acknowledgments

This work was supported in part by Zhejiang Provincial Natural Science Foundation of China under the Grant No. LZ22F020012, the National Key Research and Development Project No. 2022YFC2504605, Major Scientific Research Project of Zhejiang Lab under the Grant No. 2020ND8AD01, and in part by the Grant-in Aid for Scientific Research from the Japanese Ministry for Education, Science, Culture and Sports (MEXT) under the Grant No. 20KK0234, No. 21H03470 and No. 20K21821.

References

Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; and Wang, M. 2021. Swin-Unet: Unet-like pure Transformer for medical image segmentation. *arXiv preprint arXiv:2105.05537*.

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

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

Hu, H.; Ji, D.; Gan, W.; Bai, S.; Wu, W.; and Yan, J. 2020. Class-wise dynamic graph convolution for semantic segmentation. In *European Conference on Computer Vision*, 1–17. Springer.

Huang, H.; Cai, M.; Lin, L.; Zheng, J.; Mao, X.; Qian, X.; Peng, Z.; Zhou, J.; Iwamoto, Y.; Han, X.-H.; et al. 2021a. Graph-based pyramid global context reasoning with a saliency-aware projection for covid-19 lung infections segmentation. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1050–1054. IEEE.

Huang, H.; Xie, S.; Lin, L.; Iwamoto, Y.; Han, X.-H.; Chen, Y.-W.; and Tong, R. 2022. ScaleFormer: Revisiting the Transformer-based backbones from a scale-wise perspective for medical image segmentation. *arXiv e-prints*, arXiv–2207.

Huang, X.; Deng, Z.; Li, D.; and Yuan, X. 2021b. MissFormer: An effective medical image segmentation Transformer. *arXiv preprint arXiv:2109.07162*.

Jin, Z.; Gong, T.; Yu, D.; Chu, Q.; Wang, J.; Wang, C.; and Shao, J. 2021. Mining contextual information beyond image for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7231–7241.

Kumar, N.; Verma, R.; Sharma, S.; Bhargava, S.; Vahadane, A.; and Sethi, A. 2017. A dataset and a technique for generalized nuclear segmentation for computational pathology. *IEEE Transactions on Medical Imaging*, 36(7): 1550–1560.

Li, J.; Yu, H.; Chen, C.; Ding, M.; and Zha, S. 2022. Category guided attention network for brain tumor segmentation in MRI. *Physics in Medicine & Biology*, 67(8): 085014.

Milletari, F.; Navab, N.; and Ahmadi, S.-A. 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *Fourth International Conference on 3D Vision*, 565–571. IEEE.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, 234–241. Springer.

Schlemper, J.; Oktay, O.; Schaap, M.; Heinrich, M.; Kainz, B.; Glocker, B.; and Rueckert, D. 2019. Attention gated networks: Learning to leverage salient regions in medical images. *Medical image analysis*, 53: 197–207.

Wang, H.; Cao, P.; Wang, J.; and Zaiane, O. R. 2022a. UCTransNet: Rethinking the skip connections in U-Net from a channel-wise perspective with transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 2441–2449.

Wang, H.; Xie, S.; Lin, L.; Iwamoto, Y.; Han, X.-H.; Chen, Y.-W.; and Tong, R. 2022b. Mixed transformer U-Net for medical image segmentation. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2390–2394. IEEE.

Wang, W.; Xie, E.; Li, X.; Fan, D.-P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; and Shao, L. 2021. Pyramid vision Transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 568–578.

Woo, S.; Park, J.; Lee, J.-Y.; and Kweon, I. S. 2018. CBAM: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision*, 3–19.

Xia, Z.; Pan, X.; Song, S.; Li, L. E.; and Huang, G. 2022. Vision Transformer with deformable attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4794–4803.

Yan, X.; Tang, H.; Sun, S.; Ma, H.; Kong, D.; and Xie, X. 2022. AFTerUNet: Axial fusion Transformer U-Net for medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 3971–3981.

Yuan, Y.; Chen, X.; and Wang, J. 2020. Object-contextual representations for semantic segmentation. In *European Conference on Computer Vision*, 173–190. Springer.

Zhang, F.; Chen, Y.; Li, Z.; Hong, Z.; Liu, J.; Ma, F.; Han, J.; and Ding, E. 2019. ACFNet: Attentional class feature network for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6798–6807.

Zhou, Z.; Siddiquee, M. M. R.; Tajbakhsh, N.; and Liang, J. 2019. UNet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Transactions on Medical Imaging*, 39(6): 1856–1867.