CFII-Net: Explicit Class Embeddings and Feature Maps Through Iterative Interaction for Boosting Medical Image Segmentation

Xinyu Zhu[1] , **Xiwen Liu**[2] , **Lianghua He**[3] , **Yin Wen**[1]*
[1]East China Normal University
[2]Shanghai Ocean University
[3]Tongji University
51255904040@stu.ecnu.edu.cn, 13651858791@163.com, helianghua@tongji.edu.cn,
ywen@cs.ecnu.edu.cn

Abstract

Prior knowledge of category structure is essential in medical image segmentation, especially with significant organ structure differences. However, current hybrid architectures primarily focus on enhancing pixel-level representation learning, often neglecting or weakening the key prior knowledge of categorical structures, which poses challenges in capturing category relationships and accurate segmenting. To address this concern, we propose a novel network using Explicit Class Embeddings and Feature Maps through Iterative Interaction (CFII-Net) for boosting medical image segmentation. CFII-Net effectively segments images by exploring the relationship between explicit class embeddings and pixels in images. Specifically, we propose an Explicit Class Embedding Generator (ECEG) to obtain high-quality class semantic embeddings, incorporating category structure priors, which are used to guide high-accuracy segmentation. We then introduce an iterative Interactor, which utilizes transformers to facilitate the interaction between feature maps and class embeddings, thereby exploring pixel-to-class relationships. Furthermore, we propose updating strategies to refine the class embeddings and feature maps during the iteration process for achieving refined image segmentation. Extensive empirical evidence shows that any codec can be easily integrated into CFII-Net and yields improvements over the state-of-the-art methods in four public benchmarks.

1 Introduction

Medical image segmentation is a fundamental task in computer vision, critical for diagnosis and preoperative planning. Recently, CNN-based and Transformer-based models have achieved significant progress in this domain. A classical method U-Net [Ronneberger *et al.*, 2015] uses an encoder-decoder structure with skip connections for dense predictions. Following this technical route, its derived variants in U-Net with advanced network block techniques [Zhou *et al.*, 2018; Xiao *et al.*, 2018; Oktay *et al.*, 2018; Jin *et al.*, 2019]. Vision transformers [Dosovitskiy *et al.*, 2020] highly focus on learning the relationships between different patch tokens, which is an effective method for global context modeling. For example, TransUNet [Chen *et al.*, 2021] and SwinUNet [Liu *et al.*, 2021] both combine CNNs and transformers through serial fusion, obtaining a better trade-off in accuracy and efficiency. TransFuse [Zhang *et al.*, 2021b] introduces the BiFusion module to parallelly combine global dependencies and spatial details. However, these methods focus on obtaining better pixel representations while neglecting the structural prior knowledge of the target classes implicit in the feature maps. This can lead to semantic inconsistencies within the same class and confusion between different classes.

Recently, the mask classification segmentation paradigm has aroused the interest of many researchers in natural image segmentation tasks. This approach typically involves feeding class embeddings and feature maps into a transformer decoder to facilitate interaction, ultimately predicting a set of binary masks, each linked to a specific class. For example, SegVit [Zhang *et al.*, 2022] leverages class embeddings and attention maps to identify local patches with higher similarity, improving inference efficiency. Maskformer [Cheng *et al.*, 2021] and Mask2former [Cheng *et al.*, 2022] use the class embeddings to generate class predictions per segment, addressing semantic and instance-level segmentation tasks. Although class embeddings in the methods above are linked to category information, they are typically initialized randomly, meaning they are implicit and reliant on continuous training to learn the structural information of the categories.

In general, medical images possess significant class structural prior knowledge (e.g., organ anatomy and spatial relationships) compared to natural images, enabling the extraction of explicit and meaningful class embeddings to assist segmentation. Intuitively, shallow features capture fine-grained details, while deeper features represent abstract semantic information. In other words, each stage's features thus provide a comprehensive depiction of category-specific characteristics. Consequently, class embeddings derived from these feature maps inherit similar traits, guiding feature map regions belonging to the same category to cluster cohesively.

In light of this promising attempt, we try to extract explicit class embeddings with category structural prior in medical images and consider a simple method for predicting segmen-

*Corresponding Author

tation maps using Explicit **C**lass Embeddings and **F**eature Maps through **I**terative **I**nteraction, called CFII-Net, which consists of two parts: Explicit Class Embedding Generator (ECEG) and Iterative Interactor. Specifically, the ECEG overcomes the limitations of class-agnostic feature maps by retaining part of the original features and activating the class attributes of other features, which effectively improves the quality of class embeddings later. Subsequently, by aggregating the masks of the same class, we generate prototype class vectors to obtain explicit class embeddings. To achieve more refined class embeddings and segmentation maps, we introduce iterative interactor based on an attention mechanism. Employing this mechanism, class embeddings continuously extract information from features to improve class discriminability, and features continuously explore class prior knowledge from embeddings to enhance class relations and semantics. CFII-Net produces better results than SOTA methods with a significantly lower parameters cost (as shown in Figure 1) on synapse dataset.

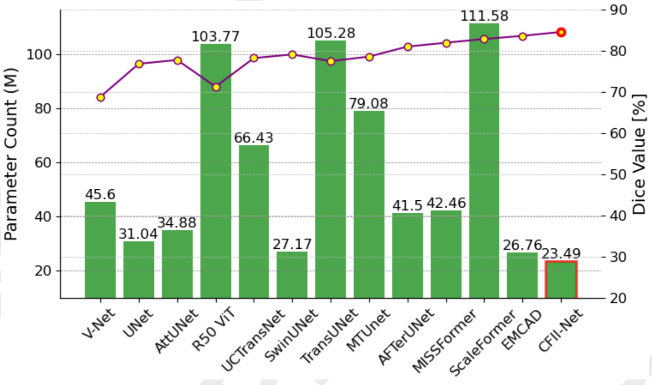


Figure 1: Parameters count vs. Dice coefficient for different methods on Synapse dataset. The green bars represent model parameters while the yellow dots show the dice coefficient for each architecture. As shown, our proposed approach (CFII-Net) has the fewest parameters, yet the highest Dice value.

Our main contributions can be summarized as follows:

- We utilize category information of feature maps to generate explicit class embeddings with structural prior knowledge to guide medical image segmentation, instead of blindly relying on random initialization.
- We build a novel iterative interaction framework CFII-Net, where the high-quality class embeddings are obtained by ECEG module, and pixel-to-class relationships between feature maps and class embeddings are explored by the Iterative Interactor, progressively refining the results.
- Extensive experimental results on four public medical datasets validate that CFII-Net outperforms state-of-the-art methods in accuracy and parameters count.

2 Related Work

2.1 CNN and Transformer for Medical Image Segmentation

Hybrid architectures based on convolutional neural network and transformer have shown excellent performance in medical image segmentation. Scaleformer [Huang *et al.*, 2022] from a scale-wise perspective to improve the segmentation quality even for small objects. UDTransNet [Wang *et al.*, 2023] explores the skip connection between encoder-decoder levels by exploring channel and spatial attention mechanisms. DTMFormer [Wang *et al.*, 2024] addresses the potential attention collapse in hybrid frameworks by proposing a plug-and-play module with dynamic token merging. EMCAD [Rahman *et al.*, 2024] proposes a new and efficient multi-scale convolutional attention decoder to obtain a better trade-off in accuracy and efficiency.

2.2 Class Embeddings for Image Segmentation

DETR [Carion *et al.*, 2020] builds an end-to-end detector by introducing learnable object queries in the transformer decoder to correspond to each object instance, which has given a lot of food for thought to image segmentation. Segmenter [Strudel *et al.*, 2021] uses a mask transformer to process image patches and class embeddings jointly and applies several self-attention layers to produce prediction maps. Maskformer [Cheng *et al.*, 2021] first realizes the importance of class embeddings and uses it to predict N class labels and N corresponding mask embeddings. Then the Mask2former [Cheng *et al.*, 2022] further advances this by introducing a new transformer decoder and employing an optimization strategy to reduce training memory. OMG-Seg [Li *et al.*, 2024] follows the Mask2Former architecture and proposes a new encoder-decoder to handle all segmentation tasks efficiently. SegViTv2 [Zhang *et al.*, 2024] transfers the similarity mapping between a set of learnable class embeddings and feature maps to the ATM modules to obtain segmentation masks. ECENet [Liu *et al.*, 2023] improves segmentation performance using roughly predicted segmentation masks for generating class embeddings.

However, since medical images contain richer category structural prior knowledge compared to natural images, randomly initialized class embeddings lead to the loss of valuable prior knowledge. Based on this, we attempt to extract meaningful explicit class embeddings from the feature maps, which are then used to iteratively interact with the features to boost medical image segmentation.

3 Method

The proposed CFII-Net framework tailored for explicit class embedding for medical image segmentation consists of two modules, i.e., 1) Explicit Class Embedding Generator (ECEG), and 2) Iterative Interactor. As shown in Figure 2, given an input image of size $H{\times}W$, we first use a backbone (such as ResUNet) to generate multi-stage features $\mathcal{F}$. Subsequently, we use $chunk$ (a simple tensor operation) on $\mathcal{F}$ along the channel dimension to form dual branches multi-stage features $\mathcal{F}_1$ and $\mathcal{F}_2$, which are used for obtaining class embeddings and refined feature maps respectively. These are defined as $\mathcal{F}_{1,i} \in \mathbb{R}^{H_i \times W_i \times C_i}$ and $\mathcal{F}_{2,i} \in \mathbb{R}^{H_i \times W_i \times C_i}$, $\forall i = 1, 2, 3, 4, 5, 6$, where $H_i = H \,/\, 2^{i-1}$ and $W_i = W \,/\, 2^{i-1}$, represents the i-th stage feature whose scale is $1/2^{i-1}$ of

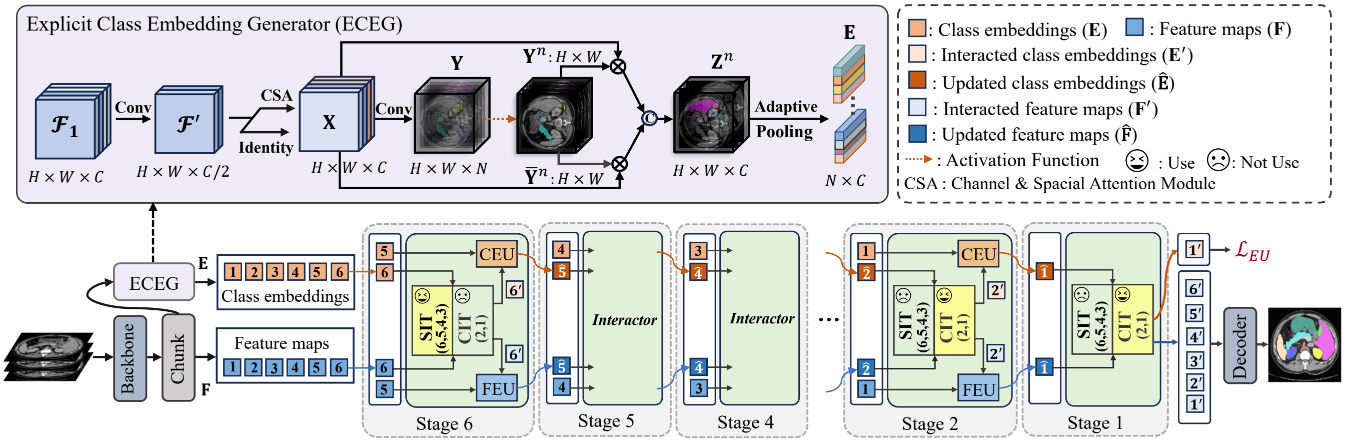


Figure 2: Overview of CFII-Net. It has two modules: (a) Explicit Class Embedding Generator (ECEG), and (b) Interactive Iterators. By extracting class prior information from feature maps into embeddings, we obtain clear and meaningful class embeddings. The illustration of the proposed Interactive Iterator module has some slight differences, deep/shallow stage features and corresponding class embeddings are sequentially fed into the SIT/CIT modules for interaction. FEU and CEU modules update feature maps and class embeddings from the previous stage, respectively. Finally, with the help of class embeddings, the updated multi-stage features are sent to the decoder.

the input image, C_i is the channel dimensions of stage i. For clarity of representation, we replace $\mathcal{F}_2$ with $\mathbf{F}$ throughout the paper. One branch $\mathcal{F}_1$ is used to generate the corresponding prototype explicit class embeddings $\mathbf{E}$ by the designed ECEG module. Then, the class embeddings and the other branch $\mathbf{F}$ are fed into multi-stages of Interactors to interact and update them iteratively. Finally, we send feature maps of each stage after interaction to the downstream block for prediction. Each module of our CFII-Net is described in detail below.

3.1 Explicit Class Embedding Generator (ECEG)

Initial multi-layer feature maps are often class agnostic, so directly mapping the feature maps to class embeddings is imprecise. To this end, we introduce a simple but effective method to obtain high-quality class semantic embeddings, dubbed Explicit Class Embedding Generator (ECEG), based on the decomposition of prediction masks.

Since the diversity and intrinsicity in features are important for obtaining high-quality category embeddings [Woo *et al.*, 2018; Han *et al.*, 2020], as observed in Figure 2, we randomly choose a branch of multi-stage features $\mathcal{F}_1$ and contend that the feature maps $\mathcal{F}_{1,i} \in \mathbb{R}^{H_i \times W_i \times C_i}$ produced in each stage contain intrinsic features. Hence, we employ linear transformations $\phi(\cdot)$ implemented by 1×1 convolutional without activation to extract these features in each stage.

$$\mathcal{F}' = \phi(\mathcal{F}_{1,i}) \in \mathbb{R}^{H_i \times W_i \times C_i/2} \tag{1}$$

To enhance the model's response to the focal points, we introduce Channel Spacial Attention (CSA) to generate attention feature maps information sequentially in both channel and spatial dimensions. These are then multiplied with the original input feature maps for adaptive feature correction, resulting in an enhanced feature maps.

$$\mathcal{F}^* = \mathcal{M}_c(\mathcal{F}') \otimes \mathcal{F}' \tag{2}$$

$$\mathcal{F}'' = \mathcal{M}_s(\mathcal{F}^*) \otimes \mathcal{F}^* \tag{3}$$

$$\mathbf{X} = [\mathcal{F}'; \mathcal{F}''] \tag{4}$$

where CSA sequentially infer a 1D channel attention map $\mathcal{M}_c \in \mathbb{R}^{C_i \times 1 \times 1}$ and a 2D spatial attention map $\mathcal{M}_s \in \mathbb{R}^{1 \times H_i \times W_i}$. $\otimes$ denotes element-wise multiplication and $[\cdot]$ denotes concatenation. Finally, the intrinsic properties branch $\mathcal{F}'$ and category properties branch $\mathcal{F}''$ are concated together and projected to the initial shape. In this way, we obtain enhanced features $\mathbf{X}$ for helping to get high-quality semantic embeddings.

Generally, convolution and attention-based operations do not explicitly decouple the category cues, so extracting category information is non-trivial. Mask classification is the assignment of a mask to each class, ideally the region in which the class is located is 1 and the background is 0. Based on this, it is logical to consider using prediction masks as the most natural raw material for extracting explicitly defined class embeddings. As shown in Figure 2, after obtaining the enhanced features $\mathbf{X}_i \in \mathbb{R}^{H_i \times W_i \times C_i}$, we first apply linear transformations to predict a coarse segmentation probability map $\mathbf{Y}_i$, which can be written as follows:

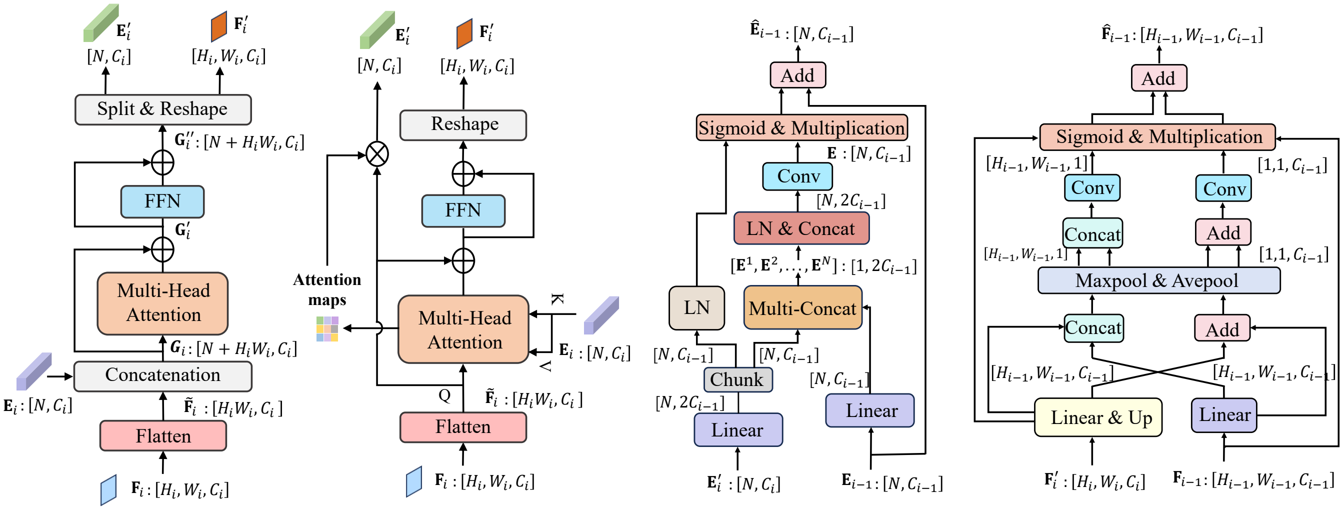
$$\mathbf{Y}_i = \phi_2(\phi_1(\mathbf{X}_i)) \in \mathbb{R}^{H_i \times W_i \times N} \tag{5}$$

where ϕ_1 and ϕ_2 are linear transformations implemented by 1×1 convolutional layers without activation, N equals to the number of classes. Then we disassemble $\mathbf{Y}_i$ by category, with each class corresponding to a class mask, and use the sigmoid function to highlight the category features and generate the category gate maps $\mathbf{Y}_i^n \in \mathbb{R}^{H_i \times W_i}$, $n \in \{1, \ldots, N\}$. We then apply both $\mathbf{Y}_i^n$ and category gate map inversion $\overline{\mathbf{Y}}_i^n$ for weighted fusion of the enhanced features $\mathbf{X}_i$.

The key insights are two folds. First, sharing the same gates can better highlight category regions. Second, the use of a subtracted gate technique supplements the missing details in the nonsalient parts. Such process is shown as follows:

$$\mathbf{Z}_i^n = \phi_3([\mathbf{X}_i \otimes \mathbf{Y}_i^n; \mathbf{X}_i \otimes \overline{\mathbf{Y}}_i^n]) \in \mathbb{R}^{H_i \times W_i \times C_i} \tag{6}$$

Finally, we aggregate the representations of all pixels, compute their average as the prototype vector for a category, and


$\mathbf{E}_i'$
$[N, C_i]$
$\mathbf{F}_i'$
$[H_i, W_i, C_i]$
Split & Reshape
$\mathbf{G}_i''$: $[N + H_iW_i, C_i]$
FFN
$\mathbf{G}_i'$
Multi-Head Attention
$\mathbf{G}_i$: $[N + H_iW_i, C_i]$
Concatenation
$\mathbf{E}_i$: $[N, C_i]$
$\tilde{\mathbf{F}}_i$: $[H_iW_i, C_i]$
Flatten
$\mathbf{F}_i$: $[H_i, W_i, C_i]$
$\mathbf{E}_i'$
$[N, C_i]$
$\mathbf{F}_i'$
$[H_i, W_i, C_i]$
Reshape
FFN
Attention maps
Multi-Head Attention
K
V
$\mathbf{E}_i$: $[N, C_i]$
Q
$\tilde{\mathbf{F}}_i$: $[H_iW_i, C_i]$
Flatten
$\mathbf{F}_i$: $[H_i, W_i, C_i]$
$\hat{\mathbf{E}}_{i-1}$: $[N, C_{i-1}]$
Add
Sigmoid & Multiplication
$\mathbf{E}$: $[N, C_{i-1}]$
Conv
$[N, 2C_{i-1}]$
LN & Concat
$[\mathbf{E}^1, \mathbf{E}^2, \ldots, \mathbf{E}^N]$: $[1, 2C_{i-1}]$
LN
Multi-Concat
$[N, C_{i-1}]$
$[N, C_{i-1}]$
Chunk
$[N, C_{i-1}]$
$[N, 2C_{i-1}]$
Linear
Linear
$\mathbf{E}_i'$: $[N, C_i]$
$\mathbf{E}_{i-1}$: $[N, C_{i-1}]$
$\hat{\mathbf{F}}_{i-1}$: $[H_{i-1}, W_{i-1}, C_{i-1}]$
Add
Sigmoid & Multiplication
$[H_{i-1}, W_{i-1}, 1]$
$[1,1, C_{i-1}]$
Conv
Conv
Concat
Add
$[H_{i-1}, W_{i-1}, 1]$
$[1,1, C_{i-1}]$
Maxpool & Avepool
Concat
Add
$[H_{i-1}, W_{i-1}, C_{i-1}]$
$[H_{i-1}, W_{i-1}, C_{i-1}]$
Linear & Up
Linear
$\mathbf{F}_i'$: $[H_i, W_i, C_i]$
$\mathbf{F}_{i-1}$: $[H_{i-1}, W_{i-1}, C_{i-1}]$

(a) Self-Interactive Transformer

(b) Cross-Interactive Transformer

(c) Class Embedding Updater

(d) Feature Enhancement Updater

Figure 3: Iterator: including interaction and updating modules.

concatenate all the category prototype vectors as explicit class embeddings $\mathbf{E}$.

$$\mathbf{E}_i^n = Pooling(\mathbf{Z}_i^n) \in \mathbb{R}^{1 \times C_i} \tag{7}$$

$$\mathbf{E} = [\mathbf{E}_i^1, \mathbf{E}_i^2; \ldots; \mathbf{E}_i^N] \in \mathbb{R}^{N \times C_i} \tag{8}$$

where $Pooling$ is the adaptive average pooling. A set of category prototype vectors is integrated by a prediction mask obtained from the enhanced feature maps, which in turn generates meaningful class embeddings. Rather than randomly initializing the class embeddings blindly, our approach explicitly extracts prior knowledge of the category structure hidden in the feature maps, which is important in the subsequent segmentation process.

3.2 Interactive Iterator

After obtaining the multi-stage explicit class embeddings $\mathbf{E}$, we feed it together with the multi-stage features $\mathbf{F}$ into the Interactive Iterator. As shown in Figure 2, we use Self-Interactive Transforme (SIT) at the deeper stage of the network and Cross-Interactive Transforme (CIT) at the shallower stage, respectively. Meanwhile, we employ Class Embeddings Updater (CEU) and Feature Enhancement Updater (FEU) to update the class embeddings and feature images progressively, which obtains more fine-grained explicit class embeddings and segmentation maps.

Self-Interactive Transformer (SIT)

As illustrated in Figure 3(a), we reshape the image features $\mathbf{F}_i \in \mathbb{R}^{H_i \times W_i \times C_i}$ into a sequence of patche tokens $\widetilde{\mathbf{F}}_i \in \mathbb{R}^{H_iW_i \times C_i}$, where $i \in \{3, 4, 5, 6\}$, then concatenate them with the class embeddings $\mathbf{E}_i \in \mathbb{R}^{N \times C_i}$ to obtain $\mathbf{G}_i \in \mathbb{R}^{(N + H_iW_i) \times C_i}$, and perform self-attention on $\mathbf{G}_i$ to ensure that the image features can explicitly recognize the class dependencies from the class embeddings. Specifically, linear transformations are applied to $\mathbf{G}_i$ to form query (Q), key (K) and value (V),

$$Q = \phi_q(\mathbf{G}_i), K = \phi_k(\mathbf{G}_i), V = \phi_v(\mathbf{G}_i) \tag{9}$$

$$\mathbf{G}_i' = Softmax(\frac{QK^T}{\sqrt{D}})V + \mathbf{G}_i \tag{10}$$

$$\mathbf{G}_i'' = FFN(\mathbf{G}_i') + \mathbf{G}_i' \tag{11}$$

where $\phi_{\alpha \in \{q,k,v\}}$ are the linear transformations, $\sqrt{D}$ serves as a scaling factor while D equals to the dimension of key. Subsequently, we pass through the feed-forward networks (FFN) and split the class embeddings and patche tokens along the channel dimension. And then reshape the patch tokens into feature maps,

$$\mathbf{E}_i', \widetilde{\mathbf{F}}_i' = Split(\mathbf{G}_i'') \tag{12}$$

$$\mathbf{F}_i' = Reshape(\widetilde{\mathbf{F}}_i') \tag{13}$$

where $\mathbf{E}_i' \in \mathbb{R}^{N \times C_i}$ and $\mathbf{F}_i' \in \mathbb{R}^{H_i \times W_i \times C_i}$ are the interacted class embeddings and interacted features obtained by reversing the order of the concatenated sequence, and are used in the iterative update in the next part.

Cross-Interactive Transformer (CIT)

The problem of quadratic complexity becomes increasingly obvious when the feature map resolution is large, so updating the feature maps sensibly is non-trivial. We aim to use class embeddings as category structural prior knowledge to help feature maps obtain clearer segmentation results, so the feature maps fed into the decoder are of utmost importance. We can effectively solve this problem by using cross-attention. As illustrated in Figure 3(b), it reshapes the image features $\mathbf{F}_i \in \mathbb{R}^{H_i \times W_i \times C_i}$ into a sequence of patche tokens $\widetilde{\mathbf{F}}_i \in \mathbb{R}^{H_iW_i \times C_i}$, where $i \in \{1, 2\}$, which are then used as a query (Q) with the class embedding as key (K) and value (V), and the rest of the operations are similar to those in SIT.

$$Q = \phi_q(\widetilde{\mathbf{F}}_i), K = \phi_k(\mathbf{E}_i), V = \phi_v(\mathbf{E}_i) \tag{14}$$

It is worth noting that, in addition to obtaining unidirectional ($\mathbf{E}\rightarrow\mathbf{F}$) interacted of the feature maps $\mathbf{F}_i'$ through the reshape operation, we also obtain interacted class embeddings $\mathbf{E}_i'$ through matrix multiplication by the mid-product attention map associated with the query.

Compared with the conventional self-attention with the complexity of $\mathcal{O}(CH^2W^2)$, the cost of cross-attention is decreased and can be summarized as $\mathcal{O}(2CNHW)$, where C is the number of channels, H,W are the height and width of the feature maps, respectively, and N is the number of segmentation categories. Even though multiple categories N are considered, the cost of cross-attention is relatively small compared to self-attention ($2N \ll HW$).

Class Embeddings Updater (CEU)

At each stage, there are unique prototype class embeddings with varying prior knowledge of class structure. Therefore, we need to continuously enrich the class embeddings.

As illustrated in Figure 3(c), after obtaining the interacted class embeddings $\mathbf{E}_i' \in \mathbb{R}^{N\times C_i}$ of the i-th layer, we refresh the previous class embeddings $\mathbf{E}_{i-1} \in \mathbb{R}^{N\times C_{i-1}}$ of the $(i-1)$-th layer through CEU. We first concatenate $\mathbf{E}_i'$ and $\mathbf{E}_{i-1}$ along the channel dimension for each category, and then apply LayerNorm (LN) [Ba *et al.*, 2016] to standardize the features of each sample:

$$\begin{aligned}\mathbf{E}^n &= \{chunk(\phi_4(\mathbf{E}_i')); \phi_5(\mathbf{E}_{i-1})\}, \mathbf{E}^n \in \mathbb{R}^{1\times 2C_{i-1}}\\ \mathbf{E} &= \phi_6([LN(\mathbf{E}^1); \ldots, LN(\mathbf{E}^N)]); \mathbf{E} \in \mathbb{R}^{N\times C_{i-1}}\end{aligned} \tag{15}$$

where $chunk(\cdot)$ indicates separation along the channel dimension. $\{\cdot\}$ represents concatenation by category, so $n \in \{1, \ldots, N\}$, N is the number of categories. Since the class features from different sources may have different distributions, this operation helps the model distinguish between different categories, thereby improving the class distinguishability. Inspired by [Zhang *et al.*, 2021a], we use a gating mechanism to refresh the previous class embeddings further. Finally, we obtain the updated class embeddings.

$$\hat{\mathbf{E}}_{i-1} = \mathit{\Psi}(\sigma(\mathbf{E}) \odot LN(chunk(\phi_4(\mathbf{E}_i'))) + \mathbf{E}_{i-1}) \tag{16}$$

where σ is the sigmoid function and $\mathit{\Psi}$ is a fully connected (FC) layer followed by LayerNorm. This iterative update gradually repeats on multi-stage features, producing high-quality class embeddings.

Feature Enhancement Updater (FEU)

Multi-scale information is particularly important for segmentation tasks. Therefore, we need to enhance the multi-scale feature information continuously. As illustrated in Figure 3(d), where the feature maps of each level are compressed to the same channel dimension by two 1 × 1 convolutional layers before entering the next level. Given an interacted feature maps $\mathbf{F}_i'$ and a previous-level feature maps $\mathbf{F}_{i-1}$, both of which have the same channel dimension, we upsample $\mathbf{F}_i'$ to the same size as $\mathbf{F}_{i-1}$ by a bilinear interpolation layer. Then, inspired by [Li *et al.*, 2019], we attempt to adapt the mechanism of automatic selection between different maps. Specifically, we first integrate the information of two branch feature maps.

$$\begin{aligned}\mathcal{A}_s &= [\phi(\mathbf{F}_i'); \phi(\mathbf{F}_{i-1})],\\ \mathcal{A}_c &= \phi(\mathbf{F}_i') + \phi(\mathbf{F}_{i-1})\end{aligned} \tag{17}$$

Then, we embed the global information through simple maximum pooling and average pooling and generate channel and spatial feature descriptors. To allow the interaction of different feature descriptors, we use a simple fully connected layer and a sigmoid activation function to obtain compact features $\mathcal{A}_{spa}$ and $\mathcal{A}_{cpa}$ for accurate and adaptive selection guidance.

$$\begin{aligned}\mathcal{A}_{spa} &= \sigma(\phi([\mathcal{P}_{max}(\mathcal{A}_s); \mathcal{P}_{avg}(\mathcal{A}_s)])),\\ \mathcal{A}_{cpa} &= \sigma(\phi(\mathcal{P}_{max}(\mathcal{A}_c) + \mathcal{P}_{max}(\mathcal{A}_c)))\end{aligned} \tag{18}$$

where $\mathcal{P}_{max}(\cdot)$ and $\mathcal{P}_{avg}(\cdot)$ are the maximum and average pooling. Finally, $\mathbf{F}_i'$ and $\mathbf{F}_{i-1}$ are matched with spatial feature descriptor and channel feature descriptor, respectively, to maximally preserve semantic information, and we get the updated outcomes $\hat{\mathbf{F}}_{i-1}$:

$$\hat{\mathbf{F}}_{i-1} = \mathcal{A}_{spa} \cdot \mathbf{F}_{i-1} + \mathcal{A}_{cpa} \cdot \mathbf{F}_i' \tag{19}$$

In brief, CEU and FEU units are reused in multi-stages, resulting in highly informative class embeddings and feature maps. The class distinguishability of the class embeddings is progressively enhanced, and valuable semantic details in the feature maps are extracted while reducing irrelevant contextual information.

Decoder and Loss

Finally, the interacted multi-level feature maps from Interactive Iterator are sent to a series of decoder blocks, including two 3×3 convolutions and a skip connection. In the training phase, we use the combined Cross-Entropy loss $\mathcal{L}_{CE}$ and Dice loss $\mathcal{L}_{Dice}$ as the primary loss. Moreover, we introduce an auxiliary Euclidean distance loss $\mathcal{L}_{EU}$ [Li *et al.*, 2022] to maximize the distance between the interacted class embeddings, which can further enhance class distinguishability.

$$\mathcal{L}_{total} = \mathcal{L}_{Dice} + \mathcal{L}_{CE} + \lambda_{EU}\mathcal{L}_{EU} \tag{20}$$

where $\mathcal{L}_{EU}$ is balanced by λ_{EU}, which is further shown in ablation experiments.

4 Experiments

4.1 Datasets

Synapse (9 classes) consists of 30 abdominal CT scans. Following [Chen *et al.*, 2021], we split 18 cases for training and 12 cases for testing. We report the Dice Coefficient (Dice) and 95% Hausdorff Distance (HD95) on 9 different organs. ACDC (4 classes) contains 100 MRI scans involving three organs. Consistent with [You *et al.*, 2022], we present the Dice results using a random split of 70 training cases, and 30 testing cases. MoNuSeg (2 classes) contains 44 images, with 30 images for training, and 14 for testing. Following [Kumar *et al.*, 2017], we report Dice and Intersection over Union (IoU) as the evaluation metrics. GlaS (2 classes) [Sirinukunwattana *et al.*, 2017] consists of 85 training samples and 80 testing samples. we report Dice as the evaluation metrics.

4.2 Implementation Details

We implement our model with PyTorch on a single NVIDIA 4090 GPU card with 24 GB of memory. To avoid overfitting, we also perform two data augmentations including random rotation and flipping. We use ResUNet as the backbone and

Method	Params (↓)	Dice (↑)	HD95 (↓)	Aorta	Gallbladder	Kidney (L)	Kidney (R)	Liver	Pancreas	Spleen	Stomach
V-Net	45.60M	68.81	-	75.34	51.87	77.10	80.75	87.84	40.05	80.56	56.98
U-Net	31.04M	76.85	39.70	89.07	69.72	77.77	68.60	93.43	53.98	86.67	75.58
AttUNet	34.88M	77.77	36.02	89.55	68.88	77.98	71.11	93.57	58.04	87.30	75.75
R50 ViT	103.77M	71.29	32.87	73.73	55.13	75.80	72.20	91.51	45.99	81.99	73.95
TransUNet	105.28M	77.48	31.69	87.23	63.13	81.87	77.02	94.08	55.86	85.08	75.62
SwinUNet	27.17M	79.12	21.55	85.47	66.53	83.28	79.61	94.29	56.58	90.66	76.60
UCTransNet	66.43M	78.23	26.75	88.86	66.97	80.19	73.18	93.17	56.22	87.84	79.43
MTUnet	79.08M	78.59	26.59	87.92	64.99	81.47	77.29	93.06	59.46	87.75	76.81
AFTerUNet	41.50M	81.02	-	**90.91**	64.81	87.90	85.30	92.20	63.54	90.99	72.48
MissFormer	42.46M	81.96	18.20	86.99	68.65	85.21	82.00	94.41	65.67	91.92	80.81
ScaleFormer	111.58M	82.86	16.81	88.73	**74.97**	86.36	83.31	95.12	64.85	89.40	80.14
EMCAD	26.76M	83.63	**15.68**	88.14	68.87	**88.08**	84.10	95.26	68.51	**92.17**	**83.92**
CFII-Net (**Ours**)	**23.49M**	**84.58**	16.59	89.26	73.66	87.19	**85.55**	**95.37**	**70.88**	91.21	83.51

Table 1: Results on Synapse dataset. The best and second-best results are **bolded** and underlined, respectively.

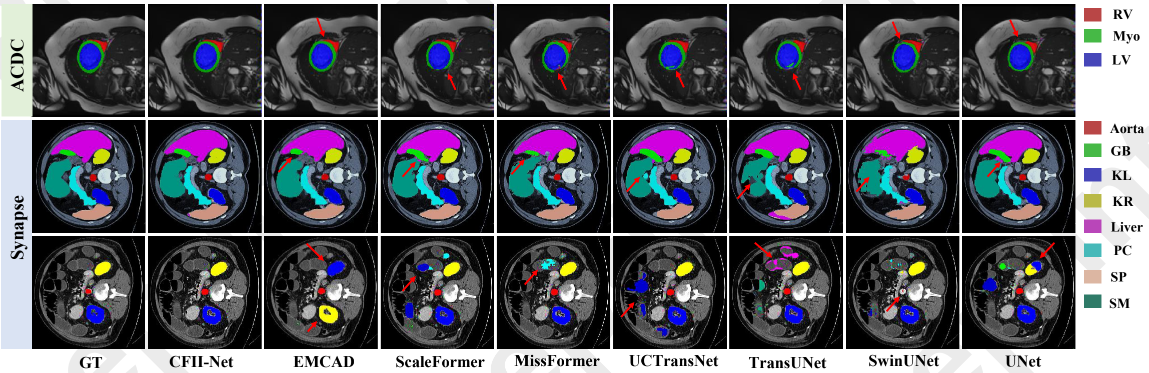


Figure 4: Qualitative results on multi-category segmentation datasets. The red arrows highlight the regions where CFII-Net performs better than the others

do not use any pre-trained weights to train the proposed CFII-Net. For all four datasets, we set the input resolution to 224 × 224 and generally train for 600 epochs.

For Synapse and ACDC, we use the SGD optimizer to train our model with a batch size of 8, where the initial learning rates are set to 0.006 and 0.04, respectively, with momentum of 0.9 and weight decay of $1e-4$. For MoNuSeg and GlaS, we train our model using the Adam optimizer, where the initial learning rates are set to 0.001 and 0.004, respectively, with a batch size of 4, and use CosineAnnealingWarmRestarts as a scheduler, with a maximum number of iterations of 10, and a minimum learning rate of $1e-4$.

4.3 Comparison with the State-of-the-Arts

Quantitative Comparison

We compare it with 12 state-of-the-art (SOTA) networks on the Synapse dataset, including V-Net [Milletari *et al.*, 2016], U-Net, AttUNet, ViT, TransUNet, SwinUNet, UCTransNet [Wang *et al.*, 2022a], MTUnet [Wang *et al.*, 2022b], AFTerUNet [Yan *et al.*, 2022], MissFormer [Huang *et al.*, 2021], ScaleFormer, and EMCAD. As shown in Table 1, although CFII-Net achieves sub-optimal results in HD95 (slightly inferior to EMCAD), it outperforms all others in the Dice metric at 84.58%. In particular, compared to ScaleFormer which only aims at extracting feature representations at different scales, we utilize the paradigm of explicit class embedding to guide the segmentation process, which maintains a better Dice metric (Dice: +1.72%) and a drastic reduction in model parameters (Params: -88.09M). Table 2 shows the Dice scores on the ACDC dataset, where CFII-Net obtains the highest average Dice score of 92.37% and outperforms other methods in all three organ segmentations.

In addition to multi-category organ segmentation, we also validate the effectiveness of CFII-Net on two binary medical image datasets. The experimental results on GlaS and MoNuSeg are reported in Table 3, where our method achieves significant advantages over others.

Method	Dice(%)	RV	Myo	LV
U-Net	87.55	87.10	80.63	94.92
AttUNet	86.75	87.58	79.20	93.47
TransUNet	89.71	88.86	84.53	95.73
SwinUNet	88.07	85.77	84.42	94.03
UCTransNet	90.42	87.28	88.54	95.44
MissFormer	90.86	89.55	88.04	94.99
ScaleFormer	90.17	87.33	88.16	95.04
EMCAD	92.12	90.65	89.68	96.02
CFII-Net (**Ours**)	**92.37**	**90.93**	**90.12**	**96.05**

Table 2: Results on the ACDC dataset.

Method	GlaS	MoNuSeg	
	Dice (%)	Dice (%)	IoU (%)
U-Net	86.76	73.97	59.48
UNet++	88.79	75.28	60.89
MedT	87.61	79.24	65.73
AttUNet	89.37	76.20	62.64
TransUNet	88.93	79.20	65.68
SwinUNet	89.67	78.49	64.72
UCTransNet	90.02	79.87	66.68
MissFormer	85.37	76.04	61.68
ScaleFormer	90.58	80.06	66.87
EMCAD	91.95	72.06	56.40
CFII-Net (**Ours**)	**92.25**	**81.33**	**68.63**

Table 3: Results on the GlaS and MoNuSeg datasets.

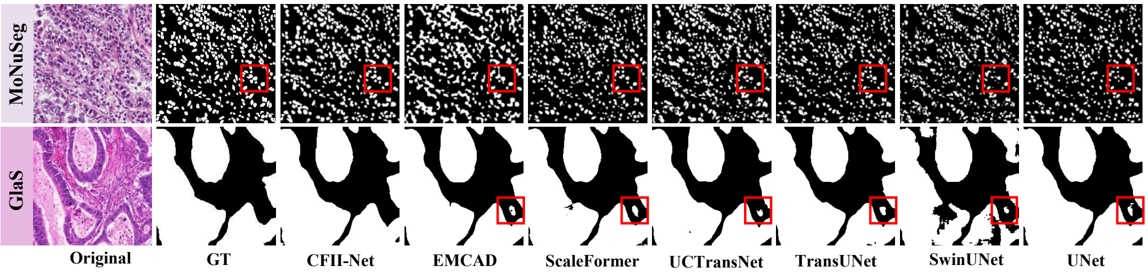


Figure 5: Qualitative results on binary segmentation datasets. The red boxes highlight areas where CFII-Net performs better than other methods.

W/CLS	FEU	CEU	Dice (%)	HD (mm)	Params (M)
✓	✓	✓	**84.58**	16.59	23.49
✓		✓	82.61	18.29	23.49
✓	✓		82.68	22.62	23.49
			78.82	33.15	44.45
✓			81.57	22.63	22.93
initialize	✓	✓	81.83	21.25	23.49

Table 4: Ablation results on the components in CFII-Net.

Backbone	W/CLS	Dice (%)	HD95 (mm)	Params (M)
UNet	×	76.85	39.70	31.04
	✓	**81.84**	**20.61**	**29.97**
TransUNet	×	77.48	31.69	105.28
	✓	**81.12**	**21.49**	**104.40**
SwinUNet	×	79.12	21.55	27.17
	✓	**80.31**	**19.74**	**26.75**

Table 5: Performance of CFII-Net on various backbones.

Qualitative Comparison

Figure 5 presents qualitative results for two binary segmentation datasets, demonstrating that CFII-Net produces outputs closer to the ground truth with fewer false segmentations. Figure 4 showcases results for multi-category segmentation datasets, highlighting the limitations of existing methods. For the ACDC dataset, EMCAD under-segments RV organs, while ScaleFormer, MissFormer, and TransUNet under-segment Myo organs. In the Synapse dataset, other methods exhibit various segmentation errors: EMCAD confuses KL and KR due to insufficient class guidance, ScaleFormer and MissFormer misidentify the background as pancreas, UCTransNet segments the background as left kidney, TransUNet misidentifies the background as liver and stomach, SwinUNet under-segments the aorta, and UNet mistakenly segments the background as gallbladder and left kidney, while confusing the right kidney with the left kidney. In contrast, CFII-Net achieves superior segmentation by leveraging explicit class embeddings enriched with structural prior knowledge. These results underscore the potential of explicit class embeddings as a promising paradigm for medical image segmentation.

4.4 Ablation Studies

Components in CFII-Net

Table 4 highlights the contributions of different components in CFII-Net. The introduction of explicit class embeddings (W/CLS), a core element of our method, significantly improves segmentation performance (Dice: +2.75%) compared to the baseline. Moreover, incorporating FEU (Dice:+1.97%) and CEU (Dice:+1.90%) modules to iteratively update feature maps and class embeddings further enhances performance with negligible additional parameters. Unlike blind randomly *initialized* class embeddings, we introduce explicit class embeddings with prior knowledge of the category structure to aid segmentation, which largely improves segmentation performance (Dice:+2.75%). These demonstrate that guided by explicitly defined class embeddings, feature maps can yield refined class cues, thus improving segmentation performance.

λ_{EU}	0	0.3	0.5	0.7	1.0
Dice (%)	84.36	**84.58**	84.48	84.53	84.38
HD (mm)	17.26	**16.59**	16.59	16.65	17.33

Table 6: Ablation results on loss coefficient.

Different Backbones

We integrate the class embeddings with three other representative medical image segmentation architectures, including CNN-based U-Net, transformer-based TransUNet, and SwinUNet, i.e., using our CFII-Net except for the backbone and decoder parts. As shown in Table 5, we find that CFII-Net achieves consistent improvements in both Dice and HD95, and also slightly reduces the model parameters.

Loss Coefficient

Finally, we evaluate the impact of different loss coefficients on performance using the Synapse dataset. The choice of loss coefficient λ_{EU} is entirely result-driven. Table 6 shows that our CFII-Net achieves an optimal 84.58% Dice when $\lambda_{EU} = 0.3$. Therefore, we use this setting in all experiments.

5 Conclusion

We present CFII-Net which uses explicit class embeddings to boost the medical image segmentation. Compared with the traditional random initialization, explicit class embedding possesses category structural prior knowledge, so regions of same category in the feature maps are more inclined to be clustered together and different categories are more distinguishable under the guidance of explicitly defined class embedding. In addition, we gradually refine feature maps and class embeddings during the iterative process, which improves the performance of the network. CFII-Net achieves state-of-the-art performance on four publicly available datasets.

Acknowledgments

This work was supported in part by the National Nature Science Foundation of China (62273150, 62171323) and the Science and Technology Commission of Shanghai Municipality under Grant 22DZ2229004.

References

[Ba *et al.*, 2016] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[Carion *et al.*, 2020] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[Chen *et al.*, 2021] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

[Cheng *et al.*, 2021] Bowen Cheng, Alexander G. Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. 2021.

[Cheng *et al.*, 2022] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022.

[Dosovitskiy *et al.*, 2020] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[Han *et al.*, 2020] Kai Han, Yunhe Wang, Qi Tian, Jianyuan Guo, Chunjing Xu, and Chang Xu. Ghostnet: More features from cheap operations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1580–1589, 2020.

[Huang *et al.*, 2021] Xiaohong Huang, Zhifang Deng, Dandan Li, and Xueguang Yuan. Missformer: An effective medical image segmentation transformer. *arXiv preprint arXiv:2109.07162*, 2021.

[Huang *et al.*, 2022] Huimin Huang, Shiao Xie, Lanfen Lin, Yutaro Iwamoto, Xianhua Han, Yen-Wei Chen, and Ruofeng Tong. Scaleformer: revisiting the transformer-based backbones from a scale-wise perspective for medical image segmentation. *arXiv preprint arXiv:2207.14552*, 2022.

[Jin *et al.*, 2019] Qiangguo Jin, Zhaopeng Meng, Tuan D Pham, Qi Chen, Leyi Wei, and Ran Su. Dunet: A deformable network for retinal vessel segmentation. *Knowledge-Based Systems*, 178:149–162, 2019.

[Kumar *et al.*, 2017] Neeraj Kumar, Ruchika Verma, Sanuj Sharma, Surabhi Bhargava, Abhishek Vahadane, and Amit Sethi. A dataset and a technique for generalized nuclear segmentation for computational pathology. *IEEE transactions on medical imaging*, 36(7):1550–1560, 2017.

[Li *et al.*, 2019] Xiang Li, Wenhai Wang, Xiaolin Hu, and Jian Yang. Selective kernel networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 510–519, 2019.

[Li *et al.*, 2022] Jiangyun Li, Hong Yu, Chen Chen, Meng Ding, and Sen Zha. Category guided attention network for brain tumor segmentation in mri. *Physics in Medicine & Biology*, 67(8):085014, 2022.

[Li *et al.*, 2024] Xiangtai Li, Haobo Yuan, Wei Li, Henghui Ding, Size Wu, Wenwei Zhang, Yining Li, Kai Chen, and Chen Change Loy. Omg-seg: Is one model good enough for all segmentation? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27948–27959, 2024.

[Liu *et al.*, 2021] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[Liu *et al.*, 2023] Yuhe Liu, Chuanjian Liu, Kai Han, Quan Tang, and Zengchang Qin. Boosting semantic segmentation from the perspective of explicit class embeddings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 821–831, 2023.

[Milletari *et al.*, 2016] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016.

[Oktay *et al.*, 2018] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.

[Rahman *et al.*, 2024] Md Mostafijur Rahman, Mustafa Munir, and Radu Marculescu. Emcad: Efficient multi-scale convolutional attention decoding for medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11769–11779, 2024.

[Ronneberger *et al.*, 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015.

[Sirinukunwattana *et al.*, 2017] Korsuk Sirinukunwattana, Josien PW Pluim, Hao Chen, Xiaojuan Qi, Pheng-Ann Heng, Yun Bo Guo, Li Yang Wang, Bogdan J

Matuszewski, Elia Bruni, Urko Sanchez, et al. Gland segmentation in colon histology images: The glas challenge contest. *Medical image analysis*, 35:489–502, 2017.

[Strudel *et al.*, 2021] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7262–7272, 2021.

[Wang *et al.*, 2022a] Haonan Wang, Peng Cao, Jiaqi Wang, and Osmar R Zaiane. 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, pages 2441–2449, 2022.

[Wang *et al.*, 2022b] Hongyi Wang, Shiao Xie, Lanfen Lin, Yutaro Iwamoto, Xian-Hua Han, Yen-Wei Chen, and Ruofeng Tong. Mixed transformer u-net for medical image segmentation. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2390–2394. IEEE, 2022.

[Wang *et al.*, 2023] Haonan Wang, Peng Cao, Xiaoli Liu, Jinzhu Yang, and Osmar Zaiane. Narrowing the semantic gaps in u-net with learnable skip connections: The case of medical image segmentation. *arXiv preprint arXiv:2312.15182*, 2023.

[Wang *et al.*, 2024] Zhehao Wang, Xian Lin, Nannan Wu, Li Yu, Kwang-Ting Cheng, and Zengqiang Yan. Dtmformer: Dynamic token merging for boosting transformer-based medical image segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5814–5822, 2024.

[Woo *et al.*, 2018] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[Xiao *et al.*, 2018] Xiao Xiao, Shen Lian, Zhiming Luo, and Shaozi Li. Weighted res-unet for high-quality retina vessel segmentation. In *2018 9th international conference on information technology in medicine and education (ITME)*, pages 327–331. IEEE, 2018.

[Yan *et al.*, 2022] Xiangyi Yan, Hao Tang, Shanlin Sun, Haoyu Ma, Deying Kong, and Xiaohui Xie. After-unet: Axial fusion transformer unet for medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 3971–3981, 2022.

[You *et al.*, 2022] Chenyu You, Ruihan Zhao, Fenglin Liu, Siyuan Dong, Sandeep Chinchali, Ufuk Topcu, Lawrence Staib, and James Duncan. Class-aware adversarial transformers for medical image segmentation. *Advances in Neural Information Processing Systems*, 35:29582–29596, 2022.

[Zhang *et al.*, 2021a] Wenwei Zhang, Jiangmiao Pang, Kai Chen, and Chen Change Loy. K-net: Towards unified image segmentation. *Advances in Neural Information Processing Systems*, 34:10326–10338, 2021.

[Zhang *et al.*, 2021b] Yundong Zhang, Huiye Liu, and Qiang Hu. Transfuse: Fusing transformers and cnns for medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021*, pages 14–24. Springer, 2021.

[Zhang *et al.*, 2022] Bowen Zhang, Zhi Tian, Quan Tang, Xiangxiang Chu, Xiaolin Wei, Chunhua Shen, et al. Segvit: Semantic segmentation with plain vision transformers. *Advances in Neural Information Processing Systems*, 35:4971–4982, 2022.

[Zhang *et al.*, 2024] Bowen Zhang, Liyang Liu, Minh Hieu Phan, Zhi Tian, Chunhua Shen, and Yifan Liu. Segvit v2: Exploring efficient and continual semantic segmentation with plain vision transformers. *International Journal of Computer Vision*, 132(4):1126–1147, 2024.

[Zhou *et al.*, 2018] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming 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*, pages 3–11. Springer, 2018.