Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.Doi Number

ViT-FRD: A Vision Transformer Model for Cardiac MRI Image Segmentation Based on Feature Recombination Distillation

Chunyu Fan[1], Qi Su[1], Zhifeng Xiao[2], Hao Su[1], Aijie Hou[1], and Bo Luan[1]
[1]Department of Cardiovascular Medicine, The People's Hospital Of Liaoning Province, Shengyang 110067 China
[2]School of Engineering, Penn State Erie, The Behrend College, Erie, 16563, PA, USA
Corresponding author: Aijie Hou (1758624242@qq.com) and Bo Luan (luanbo369@hotmail.com)

ABSTRACT Cardiac magnetic resonance imaging analysis has been a useful tool in screening patients for heart disease. Early, timely and accurate diagnosis of diseases of the heart series is the key to effective treatment. MRI provides important material for the diagnosis of cardiac diseases. The rise of deep learning has transformed computer-aided diagnostic systems, especially in the field of medical imaging. Existing work on cardiac structure segmentation models based on MRI imaging mainly relies on convolutional neural networks (CNNs), which lack model diversity and limit the prediction performance. This paper introduces Visual Transformer with Feature Recombination and Feature Distillation(ViT-FRD), a novel learning pipeline that combines a visual transformer (ViT) and a CNN through knowledge refinement. The training procedure allows the student model, i.e., ViT, to learn from the teacher model, i.e., CNN, by optimizing distillation losses. Meanwhile, ViT-FRD provides two performance boosters to increase the efficacy and efficiency of training. The proposed method is validated on two cardiac MRI image datasets. The findings demonstrate that ViT-FRD achieves SOTA and outperforms the widely used baseline model.

INDEX TERMS ViT, Feature Recombination, Feature Distillation , Heart segmentation, MRI, CNN

I. INTRODUCTION

Cardiovascular disease (CVD) is the underlying factors for 8.9 million female deaths and 9.6 million male deaths in 2019, accounting for approximately one third of all deaths globally [1]. Sometimes, even experienced radiologists may have different opinions. To this end, a computer-aided diagnosis (CAD) system can assist physicians and radiologists in the detection of heart disease via the computational and predictive power of deep learning.

Early, timely, and accurate diagnosis of heart disease is of utmost importance for effective treatment planning, disease progression monitoring, lifestyle modifications, and prevention of complications. Timely diagnosis enables healthcare professionals to develop appropriate treatment plans and closely monitor the progression of the disease. It also allows for the implementation of lifestyle modifications and preventive measures to reduce the risk of complications. In the context of cardiac foci, semantic segmentation plays a critical role by accurately localizing and delineating specific regions within the heart. This precise identification of abnormalities enables targeted interventions and quantitative assessment, enhancing the overall diagnosis and management of heart disease[2,3].

Existing cardiac MRI analysis methods face several limitations and challenges. One major limitation is the lengthy acquisition time required for cardiac MRI scans, which not only causes patient discomfort but also limits the number of patients that can be examined. Furthermore, cardiac MRI images are often affected by motion artifacts and variability due to factors like cardiac and respiratory motion, as well as magnetic field inhomogeneity. These challenges make accurate image analysis and interpretation difficult. Additionally, the complex anatomy and function of the heart present another hurdle, requiring sophisticated algorithms to segment structures, track motion, and quantify function. However, the development and adoption of such algorithms are hampered by the need for expert knowledge and manual intervention. Lastly, there is a need for better integration of cardiac MRI analysis methods with clinical decision-making, ensuring standardized protocols and automated analysis techniques that provide clinically relevant measurements for effective patient management. Addressing these limitations is crucial for enhancing the efficiency, accuracy, and clinical utility of cardiac MRI analysis[4,5].

The need for improved model diversity and prediction performance in cardiac structure segmentation is critical. Accurate segmentation of cardiac structures is essential for various clinical applications, including diagnosis and treatment planning for cardiovascular diseases. However, existing models often struggle to accurately delineate complex cardiac structures due to limited training datasets that fail to represent anatomical variations and pathologies encountered in real-world scenarios. Incorporating a more diverse and representative dataset during model training can enhance the model's ability to handle anatomical complexities and improve its overall performance. Furthermore, improving prediction performance is crucial for reliable and timely clinical information. Advancements in data acquisition and algorithm development, including comprehensive datasets and novel deep learning architectures, can enhance the models' predictive capabilities, leading to more accurate and efficient cardiac structure segmentation[6,7].

Recent advances have witnessed the prosperity of deep learning in almost all industries [5]. A wide spectrum of models, development frameworks, applications, theories, and learning paradigms have emerged and gained worldwide attention. Driven by data explosion and hardware acceleration techniques, deep neural network (DNN) models can be trained and validated with a unprecedented speed and scale, leading to a profound revolution in every industry, and health care is one of the domains that is significantly benefit from this transform. Traditional CAD systems are gradually replaced by DNN-based CAD systems, which are usually trained with massive annotated data and demonstrate superior predictive performance. DNN models have greatly improved the detection accuracy for medical imaging-based diagnoses, includes computerized tomography (CT) scans, ultrasound, MRI, and magnetic resonance imaging (MRI) .

Heart identification using MRI pictures has been investigated using a variety of DNN-based techniques [8, 9, 10, 11, 12]. Our investigation shows that most of the prior efforts utilize existing or custom convolutional neural networks (CNNs) to build heart segmentation models. These models, despite the design differences, share a core module, namely, the convolutional layer that allows the model to extract features and share parameters to reduce model size. Well-known models that have been examined include VGG , ResNet, DenseNet, EfficientNet [13], GoogleNet [14], and so on. Besides, several performance boosting techniques have been investigated, including data augmentation [15, 16], attention[17], and ensemble learning [9], which have been commonly seen in numerous DNN-based computer vision problems. Since features are extracted via convolutional layers, these CNN-based models are limited by the homogeneous internal design. It is thus necessary to explore a model with a different design principle to examine how well models other than CNNs can tackle the heart segmentation task. Transformer [18] and its variants [19, 20] are considered as promising candidates. Transformer is featured by the self-attention module, which is the core structure integrated into an encoder-decoder neural architecture. Transformer has demonstrated superior performance in numerous learning tasks, and its variant developed regarding computer vision tasks, namely, Vision Transformer (ViT) [21] has also been extensively utilized to compete against CNN models. The validated success of ViT drives us to explore its usage in heart segmentation.

Visual Transformers (ViT) are a recent development in deep learning that revolutionizes image processing tasks. Unlike Convolutional Neural Networks (CNNs), which have been the dominant architecture for computer vision, ViT applies the transformer model originally designed for natural language processing (NLP) to images. ViT breaks down an image into smaller patches, treating them as tokens similar to words in NLP. These patches are then fed into the transformer, which leverages its self-attention mechanisms to capture global relationships and dependencies across the entire image. This allows ViT to understand the context and interactions between different image regions, enabling it to learn complex patterns and make accurate predictions. By utilizing the transformer's ability to capture long-range dependencies, ViT surpasses the local receptive fields of CNNs, resulting in improved performance on various image-related tasks such as object detection, image classification, and image segmentation. The application of transformers to images has opened up exciting possibilities in computer vision and has led to significant advancements in the field.

Since ViT and CNN represent two DNN design flavors, it is desired to keep the merits of both to increase the model diversity. Knowledge distillation (KD)[22] is originally designed for model compression. KD involves two models, a teacher and a student model. The vanilla KD first trains a complex teacher model and then distills knowledge from the teacher model into a much simpler student model. We argue that the knowledge transfer within KD can not only be used for model compression, but also serves as a form of knowledge fusion between models. Knowledge, or learned patterns, can flow between the student and teacher models; meanwhile, the student model can learn patterns by itself during training. The final student can capture patterns from itself and the teacher. This finding motivates us to apply KD to heart segmentation in order for the student (ViT) to appreciate the information shared by the teacher (CNN).

Another observation is the performance gap between training and test scores, causing overfitting[23]. The phenomenon also presents in heart segmentation. The root cause of overfitting is the discrepancy of data distribution between training and test data. Traditional methods to address this issue include 1) adding more data, 2) regularization, 3) reducing the DNN’s capacity, and 4) use of dropout layers, etc. In this research article, we design a method called Cascading Cross-layer Reformer (CCReformer) to transform an image in the test set to an image that looks more similar to the training data, reducing the difference between test and training data. The following is what this paper contributes. We propose a ViT-based learning pipeline, named Knowledge-distilled and data-reformed ViT (ViT-FRD) for heart segmentation. ViT-FRD consists of three core modules, including 1) CCReformer used to reform test data during inference, 2) Linformer that allows training a ViT model with linear time and space efficiency, and 3) KD that distills knowledge from a CNN model into the ViT model. To our knowledge, this structural design has not been found in the literature. The suggested approach has shown greater performance in comparison to a number of traditional CNN-based models and the winning entries in the competition after being verified on the Multi-Modality Whole Heart Segmentation Competition 2017 (MM-WHS)dataset. The results suggest the effectiveness of the proposed method and that all three modules of the framework can boost the detection performance.

The remainder of this research articleis organized as follows. Section 2 provides a breakdown of each incorporated component as well as the overall structure of the suggested strategy. Section 3 reports the experimental details, baselines, and results. Section 4 is a paper's conclusion with a discussion of the limitations and future directions.

II. MATERIALS AND METHODS

The dataset utilized in the study is described in this part, along with specifics on the Visual Transformer with Feature Recombination and Feature Distillation（ViT-FRD）model that is suggested.

A. DATASET

The study described in this excerpt focuses on evaluating the performance of a whole-heart segmentation algorithm in the context of cardiac imaging. The validation of the algorithm is performed using two datasets: the EchoNet-Dynamic dataset and the 2017 Multimodal Whole Heart Segmentation Competition (MM-WHS) dataset.

The MM-WHS 2017 dataset [21] contains 120 multimodal cardiac images obtained in a real clinical setting and serves as the primary dataset for validating our model. The dataset comprises 20 labeled and 40 unlabeled CT volumes, as well as 20 labeled and 40 unlabeled MR volumes. Our focus in this study is specifically on MR images. The dataset is designed to evaluate various whole-heart segmentation algorithms and explore different topics related to cardiac image segmentation, registration, and modeling. As shown in Figure 1.

The EchoNet-Dynamic database [22] is a large video dataset of echocardiograms created for computer vision research in the field of cardiovascular imaging. It consists of 10,030 labeled echocardiogram videos with accompanying expert annotations, including measurements, tracings, and calculations. This dataset aims to establish a baseline for studying cardiac motion and chamber sizes using machine learning techniques. It encompasses a wide range of typical imaging conditions encountered in echocardiography laboratories. As shown in Figure 2.

Overall, the use of these two datasets allows for a comprehensive evaluation of the whole-heart segmentation algorithm and its performance in different imaging modalities and clinical scenarios.

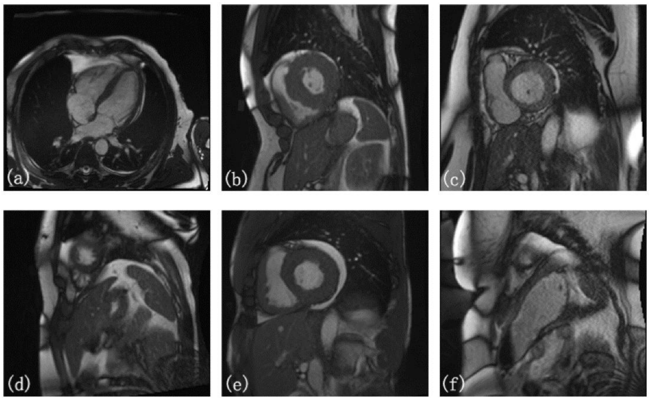

FIGURE 1. Sample heart MRI images from the MM-WHS 2017 dataset.2.2. ViT-FRD System Overview

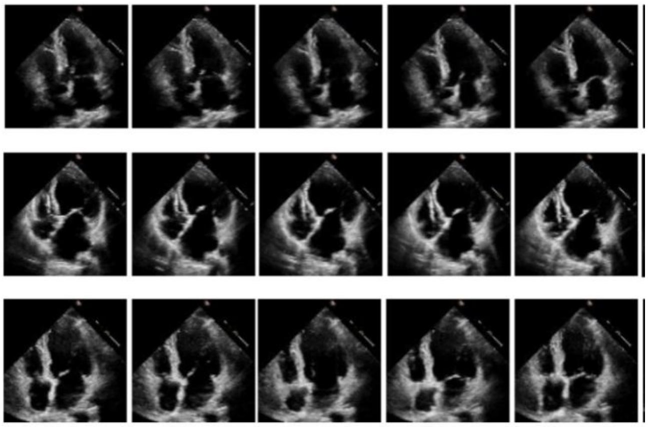

FIGURE 2. Sample cardiac ultrasound images from the EchoNet-Dynamic dataset

To provide sufficient justification for the subsequent experimental section, this paper presents the partitioning rules and results of two datasets as shown in Table 1.

The partitioning of the MM-WHS 2017 dataset is not based on group-level division, but rather involves sampling within different groups. On the other hand, for ECHONET-DYNAMIC, which comprises numerous independent dynamic videos organized into groups, no sampling is required, and the dataset can be directly partitioned by group. The specific categories for partitioning ECHONET-DYNAMIC are detailed in Table 2.

TABLE I DATA SET DIVISION RULES AND NUMBERS.

Data sets	Training	Validation	Test set	Division

	set	set		Methodology
MM-WHS 2017	80 groups	10 groups	10 groups	Stratified random sampling by image type
ECHONET-DYNAMIC	8000 groups	1000 groups	1030 groups	Stratified random sampling according to EF values

TABLE II ECHONET-DYNAMIC DIVISION DETAILS.

Data sets	Normal (EF＞=50%)	Abnormal (EF<50%)
Training set	6400 groups	1600 groups
Validation set	800 groups	200 groups
Test set	824 groups	206 groups

Figure 3 shows the proposed ViT-FRD framework for heart segmentation. The framework contains three major modules, including CCReformer, Linformer, and the KD. The CCReformer is educated through self-supervision using only the training set. During inference, a test image passes through the CCReformer, which transforms the input to an image more comparable to the ones in the drill set. After that, the transformed image is split into multiple patches that are fed into a Linformer module, which is a ViT model with linear efficiency. Meanwhile, the ViT model acts as a teacher model that has been prepared to complement a student model. The KD process allows knowledge to be transferred from the teacher to the student model by minimizing a loss function. Therefore, the student model can not only learn from the training data but also from the teacher, which has a different neural architecture and could capture patterns that the student may not see.

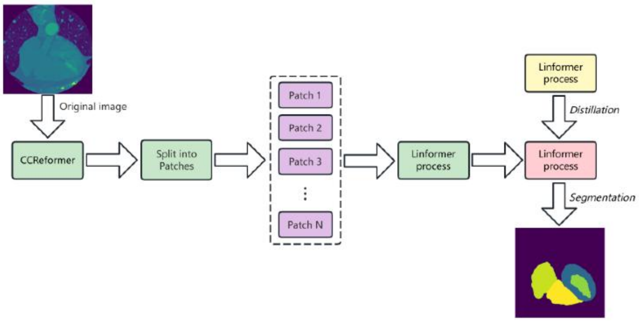

FIGURE 3. The proposed ViT-FRD learning framework. The CCReformer is trained using the training data only and used during inference to transform test data; the Linformer is a process to reduce the time complexity of ViT; the KD strategy allows the ViT student model to learn from a CNN-based teacher model for better segmentation effect.

B. TRANSFORMER AND VIT

The Transformer[18] is a neural architecture that uses attention mechanisms to process sequential data. It has several advantages over traditional recurrent models such as LSTM[31] and GRU[25], including the ability to be parallelized. The Transformer has been used for a variety of NLP applications since it was first developed for machine translation in NLP. Its success in these tasks has made it one of the most significant developments in AI in recent years.

Transformer uses a structure of encoders and decoders, in which the decoder module is made up of a stack of Transformer decoders and the encoder module is made up of a stack of Transformer decoders. For the purpose of recording the semantic relationships between the input tokens, a self-attention layer with numerous attention heads is present in each Transformer encoder. From the outputs of these attention heads, a set of embeddings is produced using the feed-forward layer, which are then fed into the next encoder.The decoder, on the other hand, has three layers: a feed-forward layer, an encoder-decoder attention layer, and a multi-head self-attention layer. At each time step, the decoder takes the embeddings from the previous step as input and produces a prediction result through layers that are linear and softmax.

Researchers are investigating Transformer's potential in computer vision as a result of its performance in NLP tasks. One example is the Vision Transformer (ViT), which modifies the original Transformer architecture to process image data. The transformation of an image into image patches, where N is the number of patches and D is the size of the patch embedding, results in a 2D tensor of size N x D. A position encoding vector is added to each patch embedding to maintain the relative position relationships between patches, and To encode information useful for categorization, a unique [CLS] token is added to the token sequence's first position. The training process for Transformer and ViT is same.

C. LINFORMER

Time-consuming nature of Transformer is O(n^2), where n refers to the input sequence length. The main efficiency bottleneck lies in the self-attention module, leading to slow speed in both training and inference, especially for inputs with a long sequence. To address this problem, Wang et al. [26] propose Linformer, a method to approximate the self-attention calculation with a low-rank matrix, reducing the complexity to O(n). Linformer has been validated to be more efficient in both time and space, while the model accuracy is on par with the standard Transformer. Therefore, we choose to integrate Linformer into the proposed ViT-FRD framework to pursue a more efficient learning pipeline.

The main idea of Linformer is to combine two matrices for linear projection to the computation of keys and values. Also, parameters are shared between projections through both the layers and the head. Specifically, there are three levels of sharing:

- Headwise sharing: Two projection matrices, E and F, are shared for each layer such that, on all heads i, Ei = E and Fi = F.
- Sharing of keys and values is subject to the aforementioned restriction. A single projection matrix E is produced for every layer.

- Layerwise sharing: The single projection matrix E is shared for all layers, all heads, keys and values in all layers.

D. CASCADING CROSS-LAYER REFORMER

By approximating the attention matrix, Linformer reduces the computational complexity of the self-attention operation to O(n), significantly improving the efficiency of the Transformer architecture, especially for inputs with long sequences. Integrating Linformer into the ViT-FRD framework allows us to effectively process image data using the Transformer architecture while addressing the efficiency problem.

The different data distribution between the training and test sets is one of the causes of the performance discrepancy between the two sets. Since the model is trained on the training set, it captures patterns based on what are presented by the training samples, which may be slightly different than the ones in the test set. This discrepancy could lead to performance degradation when evaluating the test set's model. CCReformer is suggested to solve this issue. The core idea is to train a neural network, namely, the CCReformer, with the data in the training set, so that it can transform an input image from the test set to an image that is more comparable to the ones in the training set. This way, the discrepancy between the training and test images can be eliminated, leading to a performance gain. We custom an AutoEncoder (AE) to fulfill the design requirement. An AE is usually an encoder-decoder network that goals to replicate the input to the output by reducing the loss during reconstruction, which is defined in Equation 1:

$$L_{AE} = \|\mathbf{x} - AE_{\Theta}(\mathbf{x})\|^2 \qquad (1)$$

where x refers to an input image, AEΘ is the AE network parameterized by Θ. By minimizing the L2 loss, AE is trained to capture the patterns from the input data. For our case, a custom AE is trained using all training data. A trained AE can thus transform an image in the test test to an image that is closer to the ones in the training set. We name such a trained AE as Reformer.

By stacking a series of Reformers together, we form a Cascading Cross-layer Reformer (CCReformer) module. Let K represent the number of Reformers used to construct CCReformer. The CCReformer with K layers can be defined recursively as follows.

$$CCR_1(\mathbf{x}) = RF(\mathbf{x}) \qquad (2)$$

$$CCR_2(\mathbf{x}) = RF(CCR_1(\mathbf{x}) \oplus \mathbf{x}) \qquad (3)$$

$$CCR_K(\mathbf{x}) = RF(CCR_{K-1}(\mathbf{x}) \oplus CCR_{K-2}(\mathbf{x}) \oplus \mathbf{x})) \qquad (4)$$

Where *RF* and *CCR* represent the Reformer and CCReformer modules, respectively, and $\oplus$ is the element-wise addition. Figure 4 shows an example of CCReformer (*K* = 3).

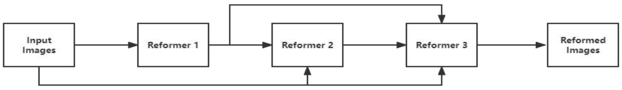

FIGURE 4. The proposed CCReformer.

E. FEATURE DISTILLATION

The main goal of using KD for this learning task is to transmit understanding from a pre-trained teacher model to our model (the student model). Since the student model is based on ViT, we tend to choose a model with a different neural architecture to be the teacher model, which is expected to learn different patterns that are otherwise may not be learned by the student model when trained individually. The KD procedure utilized in this study follows the classic setting proposed by Hinton et al. [22]. We break down the process into the following steps. First, the training set is used to train the teacher model. Second, the student model is trained with the loss function defined in Equation 5.

$$L_{KD}(\mathbf{x},y) = \alpha \cdot \|f_T(\mathbf{x}) - f_S(\mathbf{x})\|^2 + (1-\alpha) \cdot CE(y,f_S(\mathbf{x})) \qquad (5)$$

LKD consists of the distillation loss and the student loss. The former, denoted by $\|f_T(\mathbf{x}) - f_S(\mathbf{x})\|^2$ is a L2 loss, measures the soft prediction difference between the teacher and the student; the latter, denoted by $CE(y,f_S(\mathbf{x}))$, is a cross-entropy loss that represents the difference between student predictions and the ground truth. Factor α is utilized to weight the two losses.

In the process of knowledge refinement, where (x,y) represents a labeled sample from the training set, the teacher model fT and student model fS are involved. By optimizing the neural network using LKD, we ensure that the student model learns from both the instructor and the training set. This way, the trained student model, after KD, can benefit merits from both models, leading to a potential performance gain.

F. SEGMENTATION METHOD

The detection head of ViT-FRD is used to obtain the segmentation mapping by computing the outputs at different locations during the decoding phase.

In the process of decoding the ViT-FRD model, the output of each location is a fixed dimensional vector that encodes the image features at that location. To generate segmentation mappings, the ViT detection head converts these output vectors into segmentation mappings (also known as semantic segmentation maps) by feeding them into a fully connected layer that computes the segmentation probability for each location.

In the final output, the color of each pixel can be mapped to the object class to which the pixel belongs.

In summary, the detection head of ViT-FRD implements image segmentation by semantically analyzing the features at each location during the decoding process and generating segmentation mappings..

Feature Recombination and Feature Distillation (FRD) are two key components in the proposed ViT-FRD learning framework for whole-heart segmentation. Feature recombination involves combining different feature representations to capture complementary information and enhance the model's overall representation power. This is achieved through the cascading cross-layer reformer (CCReformer) module, which combines features from multiple layers using element-wise addition. By recombining features, the model can address discrepancies between the training and test sets, leading to improved generalization and performance on unseen data. Feature distillation, on the other hand, involves transferring knowledge from a pre-trained teacher model, typically a CNN, to the student model (ViT). The student model learns from the teacher model's representations and behavior, benefiting from its ability to capture patterns that the student may struggle to learn from the training data alone. Together, these techniques enhance the learning capabilities of the ViT-FRD model, improving its performance in whole-heart segmentation across different imaging modalities and clinical scenarios.

III. EXPERIMENTS AND RESULTS

This section offers details of the experiment design and implementation and reports the performance comparison results.

A. PERFORMANCE METRICS

Dice Coefficient: A measure of the similarity of two sets, usually used in semantic partitioning problems. As shown in Equation 6, where A and B represent two sets respectively, |S| denotes the number of elements in set S, and ∩ denotes the intersection.

Jaccard Coefficient (Jaccard Coefficient) or Mean Intersection-over-Union (mIoU): used in the semantic segmentation task to indicate the similarity of the predicted result to the true result. As shown in Equation 7, where A and B represent two sets respectively, |S| denotes the number of elements in set S, ∩ denotes the intersection and ∪ denotes the union.

Pixel Accuracy: indicates the proportion of all pixel points that are correctly classified in the semantic segmentation task. As shown in Equation 8, where True Positives denotes the number of correctly predicted pixels, True Negatives denotes the number of correctly predicted background pixels, and Total Pixels denotes the number of total pixels.

$$\text{Dice} = 2|A\cap B| / (|A| + |B|) \tag{6}$$

$$\text{mIoU} = |A\cap B| / |A \cup B| \tag{7}$$

$$\text{Pixel Accuracy} = (\text{True Positives} + \text{True Negatives}) / \text{Total Pixels} \tag{8}$$

The maximum Hausdorff distance (maxD) between the predicted segmentation and the ground truth segmentation. It measures the largest distance between any point in one segmentation to the nearest point in the other segmentation.

Max Hausdorff Distance is calculated as follows:

$$\text{maxD} = \max \{ \max \min \text{dist}(p, q), \max \min \text{dist}(q, p) \} \tag{9}$$

where p is a point in the predicted segmentation and q is a point in the ground truth segmentation. The function dist(p, q) represents the distance between two points.

Mean Absolute Distance (MAD): The average distance between corresponding points in the predicted segmentation and the ground truth segmentation. It provides a measure of the overall displacement between the two segmentations.

Mean Absolute Distance is calculated as follows:

$$\text{MAD} = (1/N) * \text{sum} \{ \text{dist}(p, q) \} \tag{10}$$

where N is the total number of corresponding points between the predicted segmentation and the ground truth segmentation, and dist(p, q) represents the distance between two points.

B. MODEL TRAINING

The tests are carried out using Python 3.9, while Pytorch V1.10 serves as the deep learning framework. A Windows 10 workstation with 32GB of RAM and an i7-10875h CPU was used for the experiments. A GTX2080TI graphics card was used to speed up training. When it came to training the model, we selected Adam as the optimizer and set the learning rate to 0.0001. To prevent the denominator from becoming 0, the optimizer's beta1 and beta2 parameters are set to 0.9 and 0.999, respectively, with eps=1e-08. Moreover, a batch size of eight was employed, and the weight decay was set to 0. 200 training epochs were used for all models. For the training set, we used random crop, random horizontal flip with a probability of 0.5 and resized the input images to a fixed size of 224 × 224. There is no data augmentation for the test data. The teacher model of KD was ResNet152. The AE used to build a Reformer follows a straightforward design, which consists of an encoder and a decoder. The encoder is a six-layer multi-layer perceptron (MLP) to transform a flattened image to a latent vector of size three, which then passes through a six-layer MLP decoder to recover an image of 224 × 224.

C. BASELINES

The following methods have been used as baselines for a fair comparison with the proposed method.

CNN-based models. Several representative CNN models are used as baselines, including UNet [23], FCN [24], and

SegNet [25]. Cheng Chen developed a custom model based on Transform architecture-SIFA (Synergistic Image and Feature Adaptation), which consists of two encoders with different angles and a decoder[26]. The two encoders with different angles propose a collaborative approach from the perspective of images and features. Judy Hoffman's solution is a CyCADA (Cycle-Consistent Adversarial Domain Adaptation) model based on GAN, where the authors find domain-invariant representations by using an adversarial approach and capture pixel-level and low-level domain changes and reach multiple segmented datasets in SOTA[27].

D. RESULTS AND ANALYSIS

A performance comparison between the baseline and proposed methods is reported in Table 3-4. The top half of the table presents the baseline performance, while the bottom half displays the results of the ablation study. All results are based on the two datasets mentioned in Section 2.1, and the evaluation metrics include Dice, mIoU, Pixel Accuracy, maxD and MAD, as defined in Section 3.1. The observed results are as follows:

- The proposed method outperforms the baselines, including mature CNN models like ResNet and Xception and the top-ranking solutions of the Kaggle contest in all four metrics.
- The base ViT model is worse than several CNN competitors. However, the addition of the three boosting modules have proven to be effective. Specifically, Linformer, KD, and CCReformer have brought an incremental Acc gain of 0.48%, 1.44%, and 0.48%, respectively, leading to a combined Acc gain of 2.4% compared to the base ViT model. These joint efforts allow the proposed method to be superior than the baselines as well as the SOTA.
- In addition to Acc, the other three metrics present stories about the performance gain. Specifically, Pre can be regarded as one minus the ratio of false alarms, a kind of mis-classification. A higher Pre indicates alarms. We notice that the joint gain of Pre was 1.63%, meaning that a portion of false alarms have been fixed. Also, Rec reflects the ratio of TP and all correct predictions. A higher Rec indicates that less missing for the heart cases. It is observed that a total gain of Rec was 2.3%, meaning that more heart samples that were missed previously have been detected.

TABLE III PERFORMANCE COMPARISON. ABBREVIATIONS IN THE TABLE INCLUDE LINFORMER (LIN.) AND CCREFORMER (CCR.) . on MM-WHS 2017

Frame work	Method	10%（labeled data in train set）				
		Dice	mIoU	PixAcc	maxD	MAD
CNN	U-Net	0.5647325	0.4236036	0.9735768	7.307813263	8.87914
	FCN	0.600243	0.447083	0.95195	10.94169849	12.581523
	SegNet	0.6584865	0.4985509	0.9551059	10.79153794	13.542914
	CyCADA	0.686343	0.5413044	0.95918	10.92167708	14.173181
	SIFA (SOTA)	0.7335639	0.601217	0.9538222	12.2831327	15.023388
ViT	ViT-FRD (Without Feature Recombination)	0.731	0.588	0.971	8.969589978	10.452304
	ViT-FRD (Without Feature Distillation)	0.747	0.581	0.933	11.48227646	13.184765
	ViT-FRD	**0.761**	**0.611**	**0.985**	**5.405779674**	**5.97186**

Frame work	Method	20%（labeled data in train set）				
		Dice	mIoU	PixAcc	maxD	MAD
CNN	U-Net	0.7227831	0.5818752	0.969053	15.8269216	17.729334
	FCN	0.7297774	0.60228	0.9722468	15.24630082	16.791075
	SegNet	0.7954996	0.657888	0.9626042	14.96600113	18.916235
	CyCADA	0.7964432	0.6870852	0.9619216	14.27526262	17.873484
	SIFA (SOTA)	0.8009614	0.7070944	0.971727	12.12296145	15.716358
ViT	ViT-FRD (Without Feature Recombination)	0.794	0.69	0.978	6.386828578	7.731284

	ViT-FRD (Without Feature Distillation)	0.801	**0.721**	0.966	5.986400454	7.135934
	ViT-FRD	**0.814**	0.718	**0.98**	**4.364666552**	**5.407272**

Frame work	Method	40% (labeled data in train set)				
		Dice	mIoU	PixAcc	maxD	MAD
CNN	U-Net	0.8099738	0.6951116	0.9742914	17.29849496	20.844864
	FCN	0.8209587	0.7197246	0.9812306	15.99710355	19.28786
	SegNet	0.8136853	0.7087684	0.978728	13.08398895	15.664395
	CyCADA	0.8201916	0.7158448	0.9733666	11.24881935	10.664208
	SIFA (SOTA)	0.8481696	0.7499394	0.975741	11.14764201	8.016078
ViT	ViT-FRD (Without Feature Recombination)	0.839	0.717	0.984	6.466914203	7.946446
	ViT-FRD (Without Feature Distillation)	0.823	**0.788**	0.98	4.735062566	5.517072
	ViT-FRD	**0.856**	0.763	**0.988**	**4.274570224**	**4.821257**

TABLE IV PERFORMANCE COMPARISON. ABBREVIATIONS IN THE TABLE INCLUDE LINFORMER (LIN.) AND CCREFORMER (CCR.). ON ECHONET-DYNAMIC

Frame work	Method	10% (labeled data in train set)				
		Dice	mIoU	PixAcc	maxD	MAD
CNN	U-Net	0.52362	0.4200877	0.9085419	7.942862236	9.986368758
	FCN	0.5959813	0.4349671	0.871415	11.49097175	13.26344155
	SegNet	0.649992	0.4788083	0.9132723	11.20809131	14.25256269
	CyCADA	0.6785873	0.5063903	0.8678661	12.13398324	15.94624594
	SIFA (SOTA)	0.7096497	0.5983312	0.9474316	13.90450622	16.1231
ViT	ViT-FRD (Without Feature Recombination)	0.6987629	0.566538	0.9345875	9.416275558	10.95819551
	ViT-FRD (Without Feature Distillation)	0.7374384	0.537425	0.9294546	12.56161044	14.30810698
	ViT-FRD	**0.7483674**	**0.5825885**	**0.969043**	**5.972305384**	**6.707593152**
Frame work	Method	20% (labeled data in train set)				
		Dice	mIoU	PixAcc	maxD	MAD
CNN	U-Net	0.6656832	0.5521414	0.9441483	17.67708874	19.4809922
	FCN	0.678766	0.5530135	0.9439544	16.77245553	17.72130056
	SegNet	0.7819761	0.5932834	0.8674026	16.24858743	20.90811455
	CyCADA	0.7548689	0.6557541	0.9435489	15.48723242	19.75019982
	SIFA (SOTA)	0.7351224	0.7070237	0.9483084	13.03703275	16.45031192
ViT	ViT-FRD (Without Feature Recombination)	0.7176172	0.645357	0.9078774	6.853067064	7.98255073
	ViT-FRD (Without Feature Distillation)	0.7663167	0.6678121	0.9177966	6.357557282	7.67326983
	ViT-FRD	**0.7774514**	**0.670612**	**0.952168**	**4.648369877**	**6.061551912**
Frame work	Method	40% (labeled data in train set)				
		Dice	mIoU	PixAcc	maxD	MAD
CNN	U-Net	0.7781418	0.6276858	0.9121316	18.6944835	21.79955877
	FCN	0.8123386	0.7036747	0.9619004	17.38565214	21.19928693
	SegNet	0.795296	0.6933172	0.8900552	14.55593771	16.84549038

ViT	CyCADA	0.7993587	0.6858509	0.9178847	12.60767673	11.68157344
	SIFA (SOTA)	0.8125465	0.7087677	0.9049998	12.07958489	8.585219538
	ViT-FRD (Without Feature Recombination)	0.7556034	0.6979278	0.8945544	7.271398329	8.334232565
	ViT-FRD (Without Feature Distillation)	0.7746076	0.7156616	0.902286	5.132807822	5.967265075
	ViT-FRD	**0.7852944**	**0.7577353**	**0.9572732**	**4.41434867**	**5.149584602**

IV. DISCUSSION

Current DNN-based models for cardiac MRI Image segmentation are mostly CNN models, relying on the convolutional layer for feature extraction. These models are limited by the homogeneous building blocks, affecting the predictive per-formance. Increasing model diversity is a potential strategy for performance improvement. This paper investigates a KD-based training procedure to combine the merits of two types of models, namely, ViT and CNN. KD involves two models, namely, a student (ViT) and a teacher (CNN). The training allows the student to learn knowledge from the teacher by optimizing a distillation loss. Meanwhile, the student can still learn knowledge by itself. Driven by KD, the student model appreciates the patterns discovered by both ViT and CNN, which can effectively lift the performance. On the other hand, like many other image segmentation tasks, overfitting has been an issue for heart segmentation. It is observed that the discrepancy of training and test data distribution is the key to cause the performance gap. To this end, we design CCReformer, a neural network that can learn the distribution of training data and apply to test data. The distribution of the reformed test data is closer to that of the training data, leading to a performance gain. Overall, the proposed method exhibits superior performance in four metrics when evaluated on the MM-WHS 2017 dataset, outperforming the selected CNN baseline models as well as the top-ranking solutions (i.e., SOTA) from the Kaggle competition.

The study has the following limitations that can be addressed in the future work. First, the KD procedure can involve more teacher models in addition to CNN, e.g., the Capsule Neural Network, to further enhance model diversity. In other words, the student model can learn from more than one teacher models, which may lead to further gains in performance. From the data perspective, we adopt CCReformer, an effort to close the distribution gap between training and test data. Results show that the effect of CCReformer depends on the number of reformers, namely, parameter K. It is desired to explore a general usage of CCReformer on more image segmentation tasks and develop strategies to enhance the robustness of CCReformer. Lastly, in addition to the detection result, it would be beneficial to enable the DNN models to provide evidence on the subtle difference between normal and heart samples, which would lead to a more convincing decision making process in a real world setting with better explainability.

ACKNOWLEDGMENTS

The work is not funded by any organization.

DECLARATION OF COMPETING INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data charts which are typically black and white, but sometimes include color.

REFERENCES

[1] Roth GA, Mensah GA, Fuster V. The Global Burden of Cardiovascular Diseases and Risks: A Compass for Global Action. J Am Coll Cardiol. 2020 Dec 22;76(25):2980-2981. doi: 10.1016/j.jacc.2020.11.021. PMID: 33309174. S. Dong, P. Wang, K. Abbas, A survey on deep learning and its applications, Computer Science Review 40 (2021) 100379.

[2] Muhammad Y, Tahir M, Hayat M, et al. Early and accurate detection and diagnosis of heart disease using intelligent computational model[J]. Scientific reports, 2020, 10(1): 19747.

[3] Celermajer D S, Chow C K, Marijon E, et al. Cardiovascular disease in the developing world: prevalences, patterns, and the potential of early disease detection[J]. Journal of the American College of Cardiology, 2012, 60(14): 1207-1216.

[4] Clinical cardiac MRI[M]. Springer Science & Business Media, 2012.

[5] Pirruccello J P, Bick A, Wang M, et al. Analysis of cardiac magnetic resonance imaging in 36,000 individuals yields genetic insights into dilated cardiomyopathy[J]. Nature communications, 2020, 11(1): 2254.

[6] Sander J, de Vos B D, Išgum I. Automatic segmentation with detection of local segmentation failures in cardiac MRI[J]. Scientific Reports, 2020, 10(1): 21769.

[7] Chen C, Qin C, Qiu H, et al. Deep learning for cardiac image segmentation: a review[J]. Frontiers in Cardiovascular Medicine, 2020, 7: 25.

[8] K. Simonyan, A. Zisserman, Very deep convolutional networks for large380 scale image recognition, arXiv preprint arXiv:1409.1556.

[9] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[10] G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, Densely con-nected convolutional networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4700–4708.

[11] M. Tan, Q. Le, Efficientnet: Rethinking model scaling for convolutional neural networks, in: International conference on machine learning, PMLR, 2019, pp. 6105–6114.

[12] F. Chollet, Xception: Deep learning with depthwise separable convolutions, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1251–1258.

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30.

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

[15] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis,L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretrain-ing approach, arXiv preprint arXiv:1907.11692.

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

[17] G. Hinton, O. Vinyals, J. Dean, et al., Distilling the knowledge in a neural network, arXiv preprint arXiv:1503.02531 2 (7).

[18] D. M. Hawkins, The problem of overfitting, Journal of chemical information and computer sciences 44 (1) (2004) 1–12.

[19] J. Kim, J. Kim, H. L. T. Thu, H. Kim, Long short term memory recur-rent neural network classifier for intrusion detection, in: 2016 international conference on platform technology and service (PlatCon), IEEE, 2016, pp. 1–5.

[20] J. Chung, C. Gulcehre, K. Cho, Y. Bengio, Empirical evaluation of gated recurrent neural networks on sequence modeling, arXiv preprint arXiv:1412.3555.

[21] S. Wang, B. Z. Li, M. Khabsa, H. Fang, H. Ma, Linformer: Self-attention with linear complexity, arXiv preprint arXiv:2006.04768.

[22] X. Zhuang and J. Shen, “Multi-scale patch and multimodality atlases for whole heart segmentation of mri,” Medical image analysis, vol. 31, pp. 77–87, 2016.

[23] Ouyang, D., He, B., Ghorbani, A., Lungren, M.P., Ashley, E.A., Liang, D.H., Zou, J.Y.: Echonet-dynamic: a large new cardiac motion video data resource for medical machine learning. In: NeurIPS ML4H Workshop: Vancouver, BC, Canada (2019)

[24] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. Springer International Publishing, 2015.

[25] Dai, Jifeng, et al. "R-fcn: Object detection via region-based fully convolutional networks." Advances in neural information processing systems 29 (2016).

[26] Badrinarayanan, Vijay, Alex Kendall, and Roberto Cipolla. "Segnet: A deep convolutional encoder-decoder architecture for image segmentation." IEEE transactions on pattern analysis and machine intelligence 39.12 (2017): 2481-2495.

[27] Chen, Cheng, et al. "Synergistic image and feature adaptation: Towards cross-modality domain adaptation for medical image segmentation." Proceedings of the AAAI conference on artificial intelligence. Vol. 33. No. 01. 2019.

CHUNYU FAN, He received his bachelor's degree from Bethune Clinical College of Jilin University and then received a master's degree from China Medical University. Currently, he is a lecturer at Liaoning Provincial People's Hospital. His main research interests are interventional therapy of coronary heart disease. He is good at emergency and plain clinic coronary angiography and stent implantation. He is skilled in new technologies such as coronary luminal vascular ultrasound (IVUS) and coronary flow reserve fraction measurement (FFR) and has rich experience in the treatment of perioperative complications.

QI SU, He is a young medical specialist born in 1987 and graduated from Liaoning University of Traditional Chinese Medicine with a master's degree. Currently, he is working as a lecturer in Liaoning People's Hospital.His main research interests are in the interventional treatment of coronary artery disease, and he is dedicated to the in-depth investigation of the pathological mechanism and treatment of coronary artery disease.

ZHIFENG XIAO is an Associate Professor of Department Computer Science and Software Engineering at Penn State Erie, The Behrend College. He was an Assistant Professor at Penn State Behrend from 2013 to 2019. Before that, he obtained a Ph.D. degree in Computer Science at the University of Alabama in 2013 and a B.S. degree in Computer Science at Shandong University, China, in 2008. He is broadly interested in interdisciplinary AI and cybersecurity, with a particular focus on the areas of AI-powered decision science, accountable systems, and bioinformatics.

AIJIE HOU,, She was Chief physician, second-level professor, vice president of Liaoning Provincial People's Hospital, doctoral degree, member of Zhigong, Standing Committee of Shenyang CPPCC, Special expert of The State Council, Director of Liaoning Clinical Medical Research Center for Coronary Heart Disease, Director of Cardiovascular Disease Diagnosis and Treatment Center, Director of Pharmacological Base, Director of Catheter Office. Member of the Cardiovascular Society of Chinese Medical Association, Vice Chairman of the Cardiovascular Specialty Committee of Liaoning Medical Association, Vice Chairman of the Humanities Branch of Liaoning Medical Association, master tutor of China Medical University, Dalian Medical University and Liaoning University of Traditional Chinese Medicine. He is good in interventional therapy of coronary heart disease (PCI) and chemical ablation therapy of hypertrophic obstructive cardiomyopathy, especially in radial artery coronary heart disease and interventional therapy of completely occluded vessels.

BO LUAN, He is an MD, MSc and Chief Physician with extensive medical experience and expertise. His main research interests are the application of intracavitary imaging for coronary interventions, especially for complex lesions in the coronary arteries. As the director of Cardiovascular V and the chief expert of the provincial CTO club, he has rich clinical experience and expertise, especially in the diagnosis and treatment of common diseases, multi-morbidities and difficult and serious diseases of the cardiovascular system. He has been to the University of Melbourne for training on interventional treatment of coronary artery disease and is very familiar with various advanced techniques of coronary interventional treatment, such as reverse guide wire opening CTO lesion technique, calcified lesion spin mill technique, intravascular ultrasound and OCT examination.

HAO SU, He is a medical specialist and was born in 1987. He graduated from Dalian Medical University with a master's degree and is currently working as a lecturer at Liaoning Provincial People's Hospital and is pursuing his PhD at China Medical University. His main research interests are in coronary interventional therapy, and he is dedicated to explore the treatment methods and effects of coronary heart disease in depth. Currently, he has mastered the techniques related to coronary angiography and stent implantation, and has high clinical practice experience.