UGformer for Robust Left Atrium and Scar Segmentation Across Scanners

Tianyi Liu[1], Size Hou[3], Jiayuan Zhu[2], Zilong Zhao[1], and Haochuan Jiang[1(✉)]

[1] School of Robotics, XJTLU Entrepreneur College (Taicang), Xi'an Jiaotong-Liverpool University, Taicang, Suzhou, Jiangsu 215412, P.R. China
H.Jiang@xjtlu.edu.cn
[2] School of Artificial Intelligence and Advanced Computing, XJTLU Entrepreneur College (Taicang), Xi'an Jiaotong-Liverpool University, Taicang, Suzhou, Jiangsu 215412, P.R. China
[3] School of Science, Xi'an Jiaotong-Liverpool University, SIP, Suzhou, Jiangsu 215123, P.R. China

Abstract. Thanks to the capacity for long-range dependencies and robustness to irregular shapes, vision transformers and deformable convolutions are emerging as powerful vision techniques of segmentation. Meanwhile, Graph Convolution Networks (GCN) optimize local features based on global topological relationship modeling. Particularly, they have been proved to be effective in addressing issues in medical imaging segmentation tasks including multi-domain generalization for low-quality images. In this paper, we present a novel, effective, and robust framework for medical image segmentation, namely, UGformer. It unifies novel transformer blocks, GCN bridges, and convolution decoders originating from U-Net to predict left atriums (LAs) and LA scars. We have identified two appealing findings of the proposed UGformer: 1). an enhanced transformer module with deformable convolutions to improve the blending of the transformer information with convolutional information and help predict irregular LAs and scar shapes. 2). Using a bridge incorporating GCN to further overcome the difficulty of capturing condition inconsistency across different Magnetic Resonance Images scanners with various inconsistent domain information. The proposed UGformer model exhibits outstanding ability to segment the left atrium and scar on the LAScarQS 2022 dataset, outperforming several recent state-of-the-arts.

Keywords: Left atrium segmentation · Scar prediction · Transformer · Graph convolution model

1 Introduction

Late gadolinium enhancement magnetic resonance imaging (LGE-MRI) is typically used to provide quantitative information on atrial scars [25]. In this mea-

surement, location and size in the left atrium (LA) indicate pathology (i.e., LA scars) and progression of atrial fibrillation [12].

Nowadays, deep learning models have been widely used to segment LA cavities and quantify LA scars from LGE-MRIs [3] to help radiologists with initial screening for quick pathology detection. Meanwhile, LGE-MRIs are often collected by multiple scanners and possibly in low imaging quality. Each of them produces inconsistent domain information [14], including different contrast and spatial resolutions. (1) Promoting the generalization of a segmentation model against domain inconsistency becomes another challenge (Fig. 1).
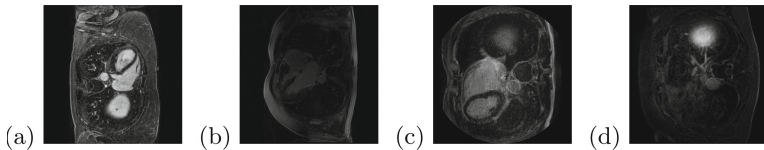

(a)   (b)   (c)   (d)

Fig. 1. Typical examples of LAScarQS Dataset [14–16] in various contrast: (a) Proper contrast, (b) low contrast, and different spatial resolution (c) 886×864, (d) 480×480.

Essentially, semantic segmentation is a mapping from input images to output pixel labels through an empirically designed segmentation model. Recent computer vision research communities have witnessed great achievements brought by the Convolutional Neural Network (CNN) and Vision Transformers (ViT) [4,10]. However, there is a lack of theoretical explanations to guarantee prediction and generalization performance [2]. Besides, there is no fixed shape in human anatomies (i.e., LAs) and pathologies (i.e., LA scars). Atlas-based segmentation strategy cannot be utilized ideally [13,30], while normal CNNs are not good at predicting deformable objects either [22].

Conventional CNN-based segmentation models only take care of local dependencies since the convolutional kernel only sees visual information in closing pixels within the receptive field. It leads to ignoring the full picture as a whole [21]. Common pooling layers in CNN will also degrade spatial information since it regards neighboring pixels as one single pixel. Losses in spatial information restrict the prediction performance of conventional CNN models [26].

Fortunately, Graph Convolutional Networks (GCN) are promised to address those challenges effectively by leveraging the robustness brought by the topological properties [11]. The topological relationship extracted by GCN while performing representation learning has been proved more stable against various application scenarios than that of the geometric relationship of general vision models, i.e., CNNs and ViTs [1]. In addition to the local features extracted by CNNs, GCN also provides an approach to model the relationship among different local features. It optimizes local features of low-quality images by Laplacian smoothing to a certain extent [9], beneficial to promoting generation across data from different domains.

Meanwhile, recently ViT models are becoming popular in semantic segmentations in handling long-range dependencies. It models spatial image information

by engaging the self-attention mechanism [24]. Swin Transformer [17] and Seg-Former [27] are two pioneering approaches to engaging ViTs in segmentation tasks. Swin Transformer engages sliding window operation. It fulfills the local-ization of convolutional operations while saving time consumption in computa-tion. SegFormer connects the transformer to lightweight multi-layer perception decoders, allowing it to combine local and global attention. In medical image segmentations, TransUnet [4], UTnet [7], and LeViT-Unet [28] are the first few trials to integrate ViT modules in the U-Net [22] architecture. All of them achieve state-of-the-art segmentation performance on the Synapse dataset [23].
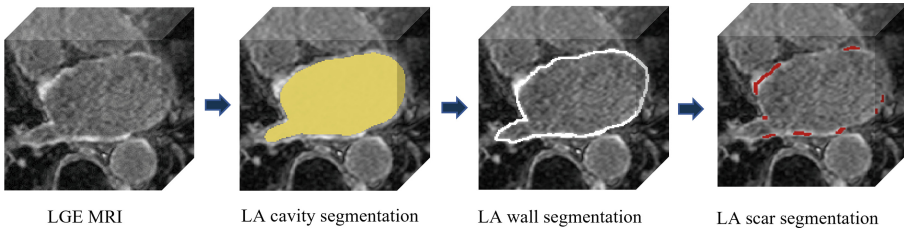

LGE MRI   LA cavity segmentation   LA wall segmentation   LA scar segmentation

Fig. 2. Positions of LA and LA scars [16]

In terms of LA scar prediction, prior work predicts LA and LA scars sepa-rately without considering the relationship between them [16]. Meanwhile, the size of the scars is relatively insignificant, bringing difficulties in the predic-tion. Fortunately, LAs are much easier to be predicted, while LA scars are often detected near identified LA boundaries Fig. 2. Inspired by [29], we believe that combining the prediction of LAs and LA scars can be expected to improve scar segmentation performance.

In this paper, we propose a novel U-shaped GCN with Enhanced Transformer module (UGformer). It is a two-stage segmentation model by segmenting the LA before quantifying the irregularly shaped LA scars. It consists of a novel transformer block as the encoder, convolution blocks as the decoder, and skip-connections with a GCN as the bridge.

In the encoder, the novel transformer block, namely, enhanced transformer block (ETB), is built by replacing the single multi-head self-attention module with paralleling the multi-head self-attention module (MHSA) and deformable convolutions (DCs). It models global spatial attention while dealing with irregu-lar shape information by leveraging advantages in both convolutions and trans-formers, i.e., proper generalization ability and sufficient model capacity [26]. The bridge with GCN connection optimizes the fusion of long-range information and context information between the encoder and the decoder [9]. It contin-uously strengthens the representation of intermediate feature maps to find a low-dimensional invariant topology, improving the extrapolation of segmenta-tion models.

The major contributions of this paper are summarized as follows:

– We proposed the UGformer, a novel two-stage segmentation model for LA and LA scar segmentation.
– In the encoder, we designed a novel enhanced transformer block combining multi-head self-attention and deformable convolutions to model global attention and address irregular shapes of LA scars.
– In the bridge, we proposed a novel GCN-based structure to optimize the global space of intermediate feature layers.
– Compared to other state-of-the-art baselines, the predicting performance of the proposed model on LAScarQS dataset [14–16] demonstrates the effectiveness and generalizability of the proposed UGformer.

2   Methodology

As depicted in Fig. 3, the proposed UGformer consists of an encoder, a U-Net decoder [22], and a bridge. Specifically, the encoder is constructed by ETB, while deconvolutions are used to build the decoder. They are connected by the bridge with GCN.
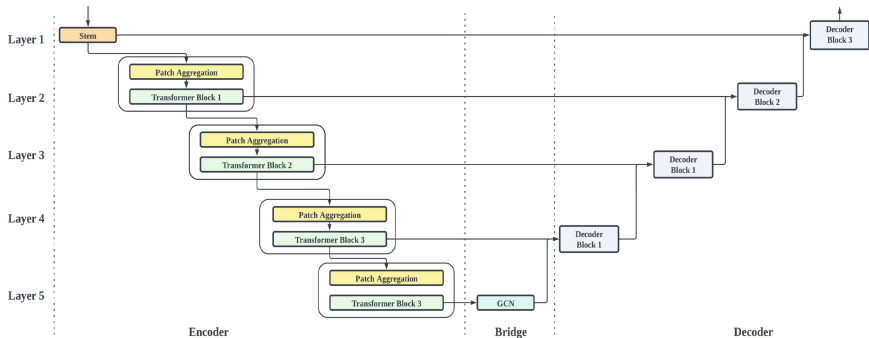

Fig. 3. UGformer Structure

2.1   Encoder Block

In the encoder, the convolutional STEM module [8], including a convolution module, a GELU module, and a batchnorm to vectorize the input features with down-sampling, was employed. It promotes quick convergence and robustness during training.

Each encoding layer (seen in Fig. 3) is constructed by a Patch Aggregation Block. Be noted that the transformer operation is not designed to downsample the feature dimension. Instead, it is constructed by the Patch Aggregation Block, including a 2×2 kernel and a stride operation with two steps to fulfill the hierarchy structure.

Besides, each layer also contains an ETB (seen in Fig. 4) to enable the UGformer to obtain both long-range dependencies and local context.
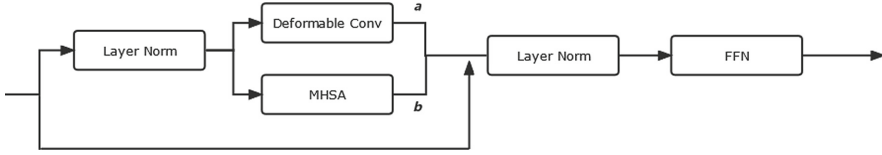

Fig. 4. EBT in UGformer

Inspired from [24], a single MHSA block is involved in ETB to extract long-range relationships and spatial dependencies. We engage DCs [5] parallel to MHSA to improve segmenting irregular LAs and quantifying LA scars. To make ETB adapt to both MHSA and deformable convolutions, a set of learnable parameters (a and b see Fig. 4) are set to leverage both paralleling parts [19].

2.2   Bridge

The bridge module is added to the skip connection from the original U-Net [22] with a GCN transformation (seen in Fig. 5). It bridges the encoder with ETB and the decoder constructed by convolutions to maximize the advantages brought by transformers and convolutions. It is capable of promoting the optimization of local features and generalization across data from different domains.
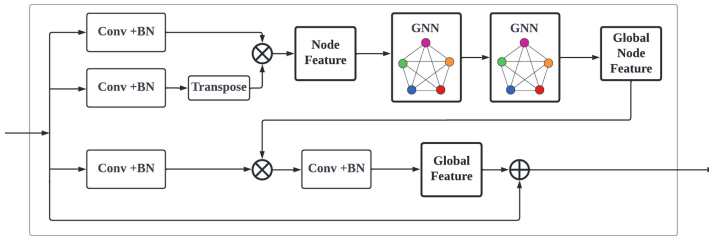

Fig. 5. The GCN Architecture in Fig. 3

GCN in Fig. 5 (see detail structure in Fig. 7) is to extract the spatial features of topological graphs by using the topologically-stable relationship information. Meanwhile, after convolutional graph operation, pixels feature belonging to the same class in semantic segmentation will be close to each other in the feature manifold (see Fig. 6).

We multiplied the feature map with the corresponding transpose as input of the GCN block. Global features will be generated by two layers of GCN blocks
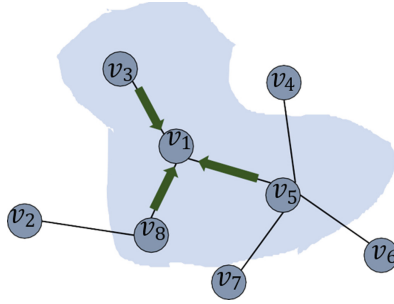
Fig. 6. GCN Topology: the global relationship of graph-based feature structure. The arrows represent the closer relationship by GCN operations in the graph. The shadow represents the topology composed of the neighbors of node v1.
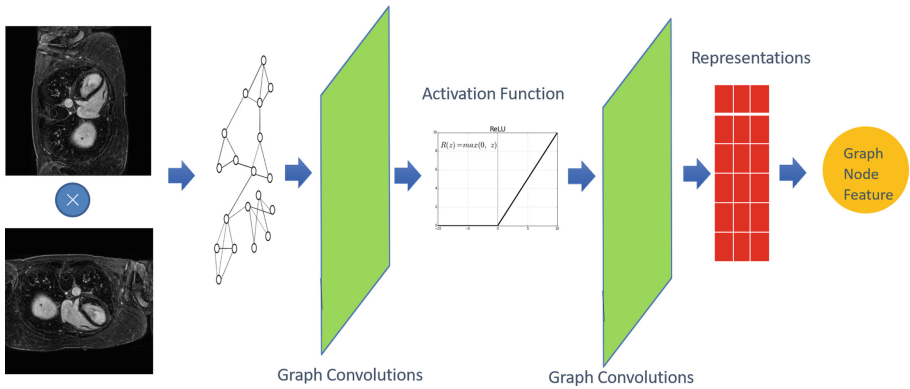

Fig. 7. Two Layers of GCN Blocks: Input feature map multiplies its transpose and update by aggregation rules in GCN block [11].

(see Fig. 7), while the global topological relationship of graph structure-based features (see Fig. 6) is obtained. The final feature map is fused by adding (see Fig. 5) the encoder output and the global relationship node feature together.

3   Implementation

3.1   Dataset and Pre-processing

The LAScarQS dataset includes two tasks: 1). LA and LA Scar segmentation (**task 1**), and 2). LA Segmentation across scanners (**task 2**). The first task contains 60 3D LGE-MRIs with labels containing LAs and LA scars, while the second consists of 130 3D LGE-MRIs from multiple medical centers with labels containing only LAs [12].
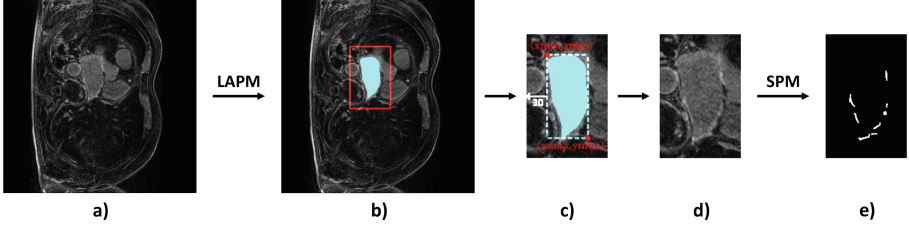
Fig. 8. Task 2 scar segmentation procedures: (a). LAMP Input, (b), Predicted LA, (c). Cropping positions, (d). Cropped ROI and SPM Input, and (e). Predicted Scar

In **task 1**, 54 subjects (approx. 44 slices per subject) are involved in the training test, while the remaining 6 subjects are used in the validation set. In **task 2**, 117 (approx. 44 slices per subject) and 13 subjects are used in the training and testing, respectively. Black margins are cropped, while images are resized to 224×224 with the bilinear interpolation before being normailzed to the range of [0, 1] by the min-max normalization. Each image is augmented 4 times by random rotation with angles sampled from $[0°, 180°]$ and translation less than $0.1 \cdot w$, where w represents the image width. The prediction performance is reported based on the 10 testing subjects available.

3.2   Training Details

We first trained the LA segmentation on **task 2**. The obtained model was loaded as the pre-training model for **task 1**. In detail, in the initial stage, the segmentation model was trained with all the LA labels available, obtaining the LA prediction model (LAPM). Then, we used the LAPM to roughly segment the targetted LA region, according to which images in the training set were cropped to train the scar prediction model (SPM). Specifically, the cropping region of interest (ROI) was implemented via $((x_{min} - 30, y_{min} - 30), (x_{max} + 30, y_{max} + 30))$, while x_{min}, x_{max}, y_{min}, y_{max} were boundary pixels of the predicted LA region, 30 was an empirically-selected tolerance of LA prediction (Fig. 8). Finally, the prediction map was restored to its original size using zero padding.

We implemented our network with the PyTorch library [20]. We ran 30 epochs on one NVIDIA Geforce RTX 3080Ti GPU. The batch size was 8, and the SGD optimizer was used. The initial learning rate was set as 10^{-4}, which would be decayed to the previous 0.1 times when the validation dice records were updated.

4   Experiment

On both tasks, we compared our UGformer with other SOTA models, including U-Net [22], Res-U-Net [6], Attention-U-Net [18]. We also performed ablation studies to demonstrate the effectiveness of our EBT and GCN bridge modules. From obtained results demonstrated in Table 1, Table 2, and Table 3, we found

that in both **task 1** and **task 2**, the proposed UGformer outperforms other baselines where transformers are engaged when evaluated by the Dice Score (DS).

4.1   Comparison to the State-of-the-art Methods (SOTA)

LA on Task 1 and Task 2: In Table 1, the dice scores outside before parentheses are performance by the model trained only with **task 1** LA dataset, while the numbers in brackets present results of models pre-trained by **task 2** dataset. We can clearly obverse that UGformer presents better prediction accuracy when predicting the LAs. Specifically, the proposed UGformer achieves the highest dice in **task 2**, outperforming all involved baselines. As shown in Fig. 9, the proposed UGformer is capable of predicting small pathological areas. At the same time, unlike Res-U-Net, UGformer is able to avoid most false detection. We believe that such an appealing factor is brought by the fact that transformers are more sensitive to irregularly shaped pathological regions [26], while the GCN module further enhances the predictive power to small regions.

We can also find from Table 1 that the Attention-U-Net performs the best no matter whether the pre-training stages are presented or not. In the meanwhile, if initialized by the pre-trained model, the DS of all the involved approaches is approx. 92 and 93. It is because that LA segmentation of **task 1** is a relatively simple assignment with consistent style information since they are generated from one single scanner.

Scar on Task 1: The proposed UGformer performs the best in this scenario by at least 2.5% compared to other baselines. It demonstrates that it is particularly useful in quantifying irregular and scattered LA scars. As shown in Fig. 10, UGformer clearly identifies more pathological regions and contributes to fewer false detections.

Table 1. Comparison between SOTA models.

Method	Task 1-LA	Task 1-Scar	Task 2-LA
	DS↑	DS↑	DS↑
U-Net	85.95 (92.24)	67.76	84.42
Res-U-Net	85.26 (92.28)	62.61	83.74
Attention-U-Net	87.40 (**93.22**)	70.11	85.37
UGformer	85.49 (92.36)	**72.66**	**86.59**

4.2   Ablation Studies

Influence of ETB Module: In Table 2, ablations of MHSAs and DCs in the ETB are presented. We can conclude that both MHSAs and DCs are essential to

<table>
<tr><td>(a) Image</td><td>(b) Gound Truth</td><td>(c) U-Net</td></tr>
<tr><td>(d) Res-U-Net</td><td>(e) Attention-U-Net</td><td>(f) UGformer</td></tr>
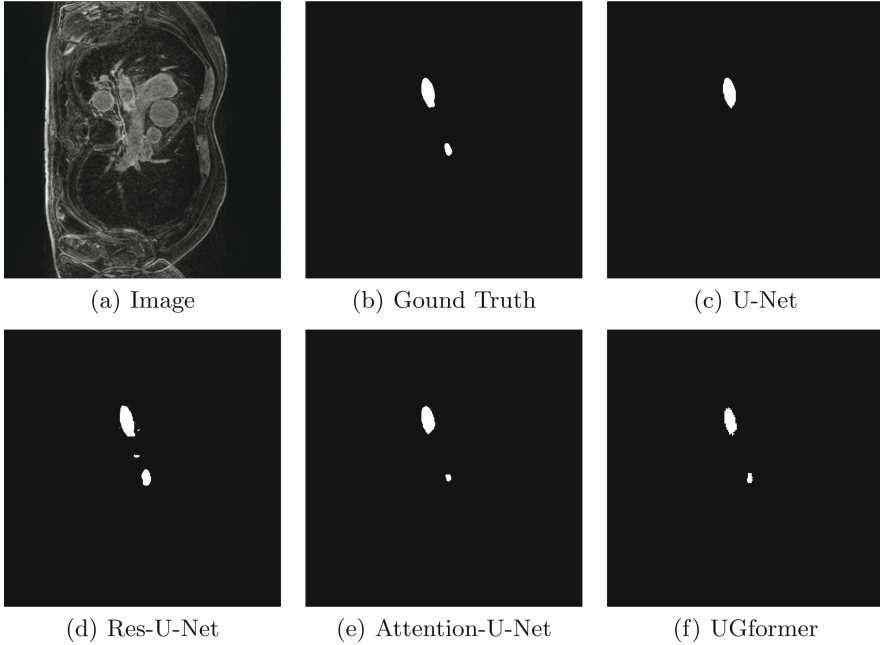</table>

Fig. 9. Prediction results on task 2 LA.

achieve the best segmentation performance at 85.49%, 72.66%, and 86.59% on DS on **task 1-LA**, **task 1-Scar**, and **task 2-LA**, respectively. Particularly, the combination of MHSAs and DCs module makes the greatest significant improvement on **task 2-LA** by 7%. It proves that the two modules contribute to each other and help the prediction of the model.

Table 2. Comparison of ETB module.

MHSA	DC	Task 1-LA	Task 1-Scar	Task 2-LA
		DS	DS	DS
✓		85.06	69.65	78.67
	✓	85.26	70.50	80.66
✓	✓	**85.49**	**72.66**	**86.59**

Influence of GCN: Table 3 enumerates the results of ablations of GCN block when the proposed UGformer and U-Net are used as backbones. From there, we can find that GCN improves the prediction performance of U-Net in **task 1-LA** and **task 2-LA**. However, the improvement in scar prediction in **task 1-Scar** with U-Net is insignificant. When GCN is implemented in the UGformer
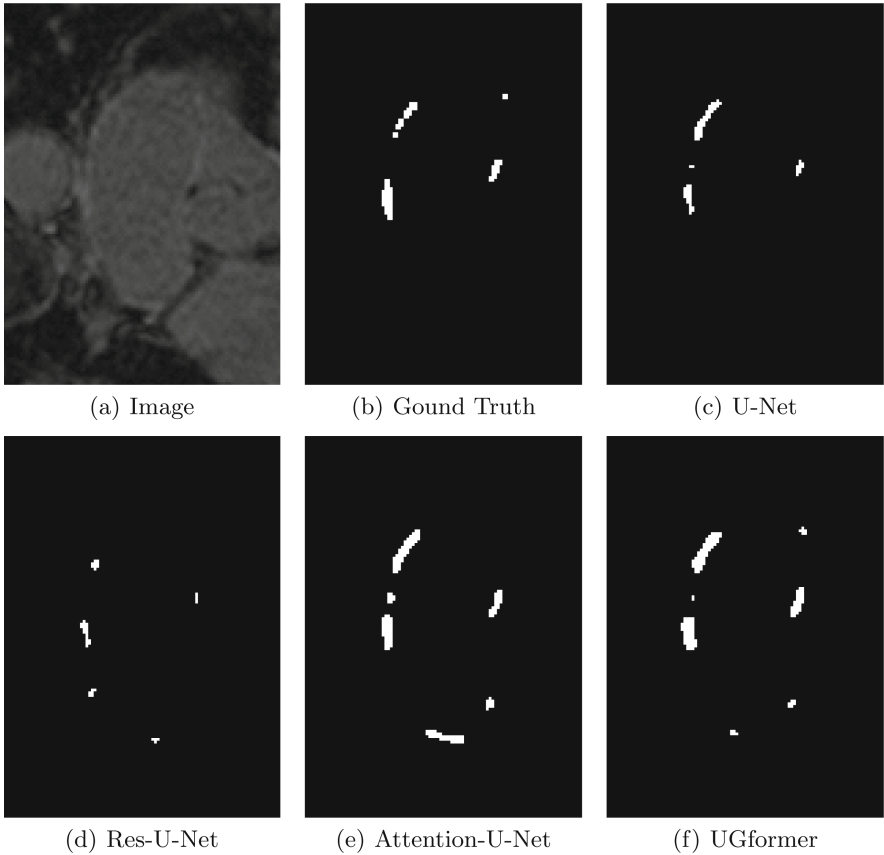
(a) Image   (b) Gound Truth   (c) U-Net

(d) Res-U-Net   (e) Attention-U-Net   (f) UGformer

Fig. 10. Prediction results on task 1 Scar. Res-U-Net can not predict the pathology. U-Net and Attention-U-Net can predict a certain part of the pathology. Nevertheless, we can also obverse worse false detection than that predicted by the proposed UGformer.

architecture, it improves the prediction performance in all settings. Particularly, when predicting scars, GCN module improves the transformer performance from 70.82% to 72.66% by 2.6%.

Table 3. Comparison of different bridge module.

Architecture	GCN	Task 1-LA DS	Task 1-Scar DS	Task 2-LA DS
U-Net		85.95	**67.76**	84.42
	✓	**87.93**	67.72	**86.79**
UGformer		84.47	70.82	85.44
	✓	**85.49**	**72.66**	**86.59**

Influence of the Two-Stage Method: Figure 11 displays the prediction results with the two-stage prediction approaches and the normal ones. It can be clearly seen that the two stage method has successfully predicted most of the scars (see Fig. 11(c)), although some kind of false detection can still be observed. Nevertheless, with the common prediction method (see Fig. 11(f)), the scar is almost impossible to be predicted. We can hereby conclude that the two-stage prediction approach is essential in quantifying scars with irregular and tiny occupations on the picture.
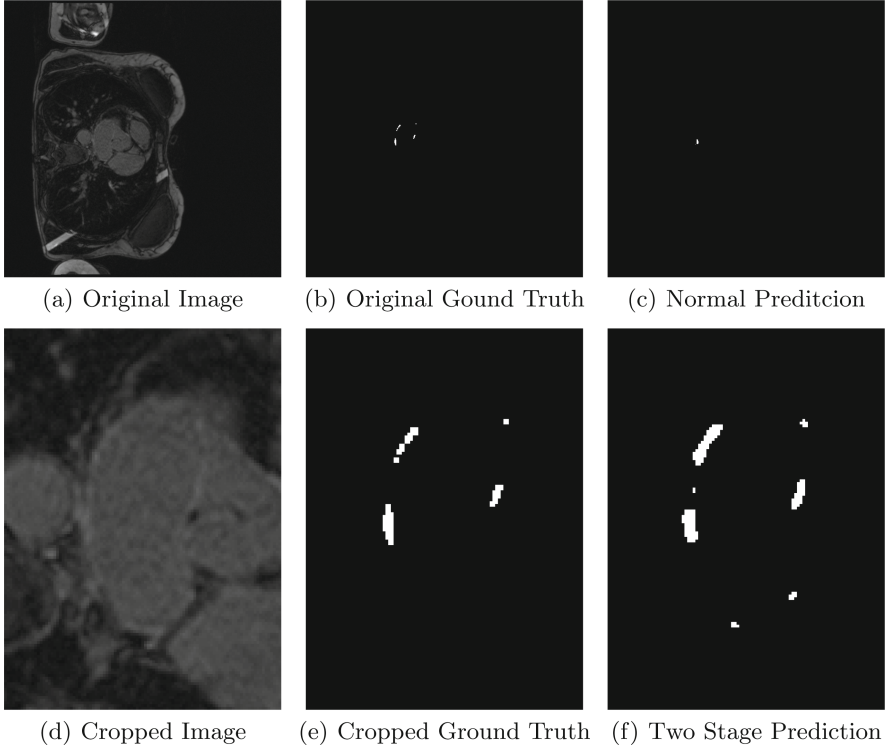

(a) Original Image   (b) Original Gound Truth   (c) Normal Preditcion

(d) Cropped Image   (e) Cropped Ground Truth   (f) Two Stage Prediction

Fig. 11. Prediction results on original images and cropped images

5 Conclusions

In this paper, we proposed the UGformer, a novel U-shaped transformer architecture with a GCN bridge. It is capable of segmenting the left atrium (LA) across different scanners and quantifying LA scars with a two-stage predicting strategy given late gadolinium enhancement magnetic resonance images. Specifically, an enhanced transformer block combining multi-head self-attention and deformable

convolutions is introduced to model global attention and overcome degradation in quantifying scars with irregular shapes. We also employ a graph convolution network (GCN), a novel GCN-based bridge, to optimize the global space of intermediate feature layers. Extensive empirical experiments on the LAScarQS 2022 challenge dataset have demonstrated the effectiveness and robustness of the proposed UGformer architecture in LA prediction and scar quantification.

References

1. Carlsson, G., Gabrielsson, R.B.: Topological approaches to deep learning. In: Baas, N.A., Carlsson, G.E., Quick, G., Szymik, M., Thaule, M. (eds.) Topological Data Analysis. AS, vol. 15, pp. 119–146. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-43408-3_5

2. Chan, K.H.R., Yu, Y., You, C., Qi, H., Wright, J., Ma, Y.: Redunet: a white-box deep network from the principle of maximizing rate reduction. J. Mach. Learn. Res. **23**(114), 1–103 (2022). http://jmlr.org/papers/v23/21-0631.html

3. Chen, C., Bai, W., Rueckert, D.: Multi-task learning for left atrial segmentation on GE-MRI. In: Pop, M., Sermesant, M., Zhao, J., Li, S., McLeod, K., Young, A., Rhode, K., Mansi, T. (eds.) STACOM 2018. LNCS, vol. 11395, pp. 292–301. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-12029-0_32

4. Chen, J., et al.: Transunet: transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)

5. Dai, J., et al.: Deformable convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 764–773 (2017)

6. Diakogiannis, F.I., Waldner, F., Caccetta, P., Wu, C.: Resunet-a: a deep learning framework for semantic segmentation of remotely sensed data. ISPRS J. Photogramm. Remote. Sens. **162**, 94–114 (2020)

7. Gao, Y., Zhou, M., Metaxas, D.N.: UTNet: a hybrid transformer architecture for medical image segmentation. In: de Bruijne, M., Cattin, P.C., Cotin, S., Padoy, N., Speidel, S., Zheng, Y., Essert, C. (eds.) MICCAI 2021. LNCS, vol. 12903, pp. 61–71. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87199-4_6

8. Guo, J., Han, K., Wu, H., Tang, Y., Chen, X., Wang, Y., Xu, C.: Cmt: convolutional neural networks meet vision transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12175–12185 (2022)

9. Han, K., Wang, Y., Guo, J., Tang, Y., Wu, E.: Vision gnn: an image is worth graph of nodes. arXiv preprint arXiv:2206.00272 (2022)

10. Huang, X., Deng, Z., Li, D., Yuan, X.: Missformer: an effective medical image segmentation transformer. arXiv preprint arXiv:2109.07162 (2021)

11. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016)

12. Lab, F.Z.: Lascarqs 2022: Left atrial and scar quantification & segmentation challenge. [EB/OL]. https://zmic.fudan.edu.cn/lascarqs22/. Accessed June 30, 2022

13. Li, L., Wu, F., Yang, G., Xu, L., Wong, T., Mohiaddin, R., Firmin, D., Keegan, J., Zhuang, X.: Atrial scar quantification via multi-scale cnn in the graph-cuts framework. Med. Image Anal. **60**, 101595 (2020)

14. Li, L., Zimmer, V.A., Schnabel, J.A., Zhuang, X.: AtrialGeneral: domain generalization for left atrial segmentation of multi-center LGE MRIs. In: de Bruijne, M., Cattin, P.C., Cotin, S., Padoy, N., Speidel, S., Zheng, Y., Essert, C. (eds.) MICCAI 2021. LNCS, vol. 12906, pp. 557–566. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87231-1_54

15. Li, L., Zimmer, V.A., Schnabel, J.A., Zhuang, X.: Atrialjsqnet: a new framework for joint segmentation and quantification of left atrium and scars incorporating spatial and shape information. Med. Image Anal. **76**, 102303 (2022)
16. Li, L., Zimmer, V.A., Schnabel, J.A., Zhuang, X.: Medical image analysis on left atrial lge mri for atrial fibrillation studies: a review. Medical Image Analysis, p. 102360 (2022)
17. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (2021)
18. Oktay, O., et al.: Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999 (2018)
19. Pan, X., et al.: On the integration of self-attention and convolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 815–825 (2022)
20. Paszke, A., et al.: Pytorch: an imperative style, high-performance deep learning library. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems 32, pp. 8024–8035. Curran Associates, Inc. (2019). http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf
21. Raghu, M., Unterthiner, T., Kornblith, S., Zhang, C., Dosovitskiy, A.: Do vision transformers see like convolutional neural networks? Adv. Neural. Inf. Process. Syst. **34**, 12116–12128 (2021)
22. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
23. Synapse, M.: Multi-atlas labeling beyond the cranial vault - workshop and challenge. [EB/OL]. https://www.synapse.org/#!Synapse:syn3193805/wiki/89480. Accessed June 30, 2022
24. Vaswani, A., et al.: Attention is all you need. Advances in neural information processing systems 30 (2017)
25. Vergara, G.R., Marrouche, N.F.: Tailored management of atrial fibrillation using a lge-mri based model: from the clinic to the electrophysiology laboratory. J. Cardiovasc. Electrophysiol. **22**(4), 481–487 (2011)
26. Xiao, T., Singh, M., Mintun, E., Darrell, T., Dollár, P., Girshick, R.: Early convolutions help transformers see better. Adv. Neural. Inf. Process. Syst. **34**, 30392–30400 (2021)
27. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: simple and efficient design for semantic segmentation with transformers. Adv. Neural. Inf. Process. Syst. **34**, 12077–12090 (2021)
28. Xu, G., Wu, X., Zhang, X., He, X.: Levit-unet: make faster encoders with transformer for medical image segmentation. arXiv preprint arXiv:2107.08623 (2021)
29. Zhou, X., Koltun, V., Krähenbühl, P.: Probabilistic two-stage detection. arXiv preprint arXiv:2103.07461 (2021)
30. Zhu, L., Gao, Y., Yezzi, A., Tannenbaum, A.: Automatic segmentation of the left atrium from mr images via variational region growing with a moments-based shape prior. IEEE Trans. Image Process. **22**(12), 5111–5122 (2013)