SANGRE: a Shallow Attention Network Guided by Resolution Expansion for MR Image Segmentation

Ying He[1,2], Marc E. Miquel[2,3,4], and Qianni Zhang[1,2,3]

[1] School of Electronic Engineering and Computer Science, Queen Mary University of London, London, UK
ying.he@qmul.ac.uk, qianni.zhang@qmul.ac.uk
[2] Clinical Physics, Barts Health NHS Trust, London, UK
[3] Digital Research Institute, Queen Mary University of London, London, UK
[4] Advanced Cardiovascular Imaging, William Harvey Research Institute, Queen Mary University, London, UK

Abstract. Magnetic Resonance (MR) imaging plays a vital role in clinical diagnostics and treatment planning, with the accurate segmentation of MR images being of paramount importance. Vision transformers have demonstrated remarkable success in medical image segmentation; however, they fall short in capturing the local context. While images of larger sizes provide broad contextual information, such as shape and texture, training deep learning models on such large images demands additional computational resources. To overcome these challenges, we introduce a shallow attention feature aggregation (SAFA) module to progressively enhance features' local context and filter out redundant features. Moreover, we use feature interactions in a resolution expansion guidance (REG) module to leverage the wide contextual information from the images at higher resolution, ensuring adequate exploitation of small class features, leading to a more accurate segmentation without a significant increase in FLOPs. The model is evaluated on two dynamic MR datasets for speech and cardiac cases. The proposed model outperforms other state-of-the-art methods. The codes are available at https://github.com/Yhe9718/SANGRE.

Keywords: MRI segmentation · Shallow Attention · Resolution Expansion

1 Introduction

Medical image segmentation is crucial for assisting clinicians in formulating treatment plans and evaluating post-treatment conditions of various diseases. The segmentation task involves classifying images at the pixel level to generate maps that can delineate relevant structures. Recently, deep learning methods such as UNet have advanced the state-of-the-art (SOTA) in this area [1][2][3]. UNet is a neural network with an encoder-decoder structure, which was first

designed for medical image segmentation and has subsequently been applied in a wide range of scenarios. Among the variants of UNet, UNet++[4], UNet3+[5] and DC-UNet[3] have demonstrated promising performance for medical image segmentation. However, in various tasks, CNN-based models have shown their limitations in modelling the long-range spatial dependencies within an image. To overcome this limitation, the attention mechanism has been developed and incorporated into the CNN models [2]. The attention mechanism improves the segmentation performance by selecting only a subset of important features to detect the targeted objects. The vision transformer focuses on modelling long-range dependencies with self-attention that capture correlations between all input tokens [6]. The vision transformer (ViT) was first proposed by Dosovitskiy et al. for the classification task[6]. It splits an image into non-overlapping patches, which are then fed into the transformer layer with positional embedding. A multi-head self-attention module is utilised to capture the long-range dependencies. Meanwhile, Liu et al. introduced the Swin Transformer, enhancing computing efficiency through shifting windows-based attention [7]. Similar to the Swin Transformer, the Pyramid Vision Transformer (PVT) is another hierarchical vision transformer [8], utilizing spatial reduction attention to improve computational efficiency. Encouraged by the success of vision transformers in various computer vision tasks, several transformer-based segmentation networks have been introduced, marking a further leap forward in medical image segmentation [9][10], e.g., TransUNet was proposed to segment CT images [11].

Despite the recent progress in image segmentation introduced by CNNs and transformers, both methods have limitations. The locality of convolutional operation limits CNN's scope to capture the global context, while the transformer suffers from its constrained localization abilities, stemming from the inadequate low-level feature representation. To integrate the advantages of both models while overcoming their limitations, numerous hybrid models have been proposed. Feiniu et al. use dual encoder consisting of a convolutional neural network branch and a transformer branch to encode the images and produce complementary features [12], while Yundong et al. utilize two CNN decoders to capture global dependency and low-level spatial details, and then fuse the multi-scale features of the two branches with a fusion module [13]. PVT-CASCADE relies on a pyramid vision transformer to extract multi-scale features and a decoder that progressively refines the encoded features while enhancing the long-range and local context using attention modules [14].

In this study, we propose a new Shallow Attention Network Guided by Resolution Expansion (SANGRE). It starts with a transformer encoder, which produces encoded features that are enhanced by a shallow attention feature aggregation (SAFA) module and a resolution expansion guidance (REG) module. The SAFA module progressively enriches and aggregates the features with local context, while REG module exploits contextual information from images with expanded resolution to refine the feature map by interacting features of expanded-resolution and coarse feature maps. By effectively leveraging both local

and global contexts, the model excels in identifying small objects with enhanced precision in shape details.

The key contributions of this study include: (1) The SAFA module allows the local context to be progressively integrated to enrich the features. (2) The REG module allows the broad context of an image with expanded resolution to be used as prior knowledge to refine the coarse feature map. (3) The features on multiple scales are progressively aggregated to complement the shortcomings of Transformers by involving both local and broad context.

2 Method

The proposed Shallow Attention Network Guided by Resolution Expansion (SANGRE) is schematically depicted in Figure 1. The network entails three branches: (1) a transformer encoder that extracts multi-scale features with global context from the training images; (2) a shallow attention feature aggregation (SAFA) module to effectively fuse the multi-scale features for local context enhancement in an accumulative manner; and (3) a resolution expansion guidance (REG) module that utilizes the broadened contextual information from images with expanded resolution, to refine the feature map from SAFA.

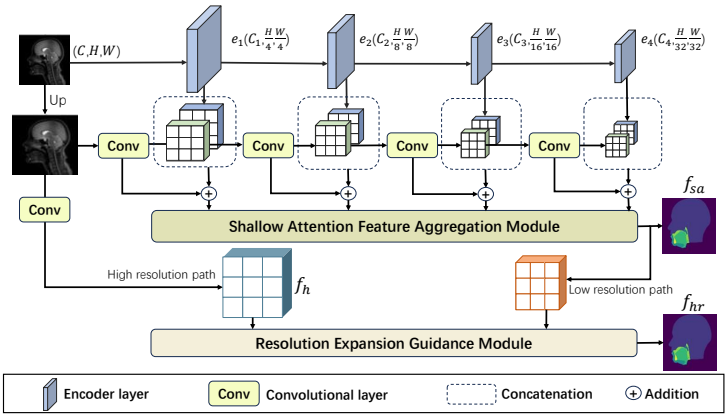

Fig. 1. Architecture of the proposed SANGRE network.

2.1 Transformer encoder

The proposed network first employs the Pyramid vision transformer(PVTv2) to extract features from the images [15]. For an input image I of size C, H, W to the network, four feature maps e_i, for i = 1,2,3,4 with size of $C_i, \frac{H}{2^{i+1}}, \frac{W}{2^{i+1}}$, are extracted by the PVTv2, where C_i is the channel number of e_i. The PVTv2 is pre-trained on ImageNet [16], to exploit the visual knowledge, which ease the overfitting problem common to transformer encoders.

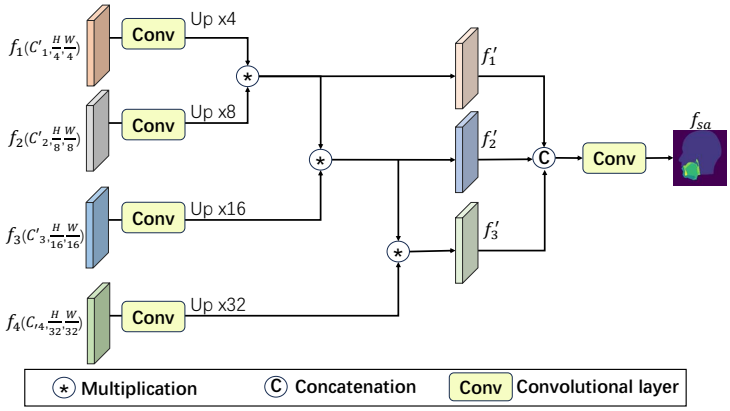

Fig. 2. Schematic diagram of the SAFA module.

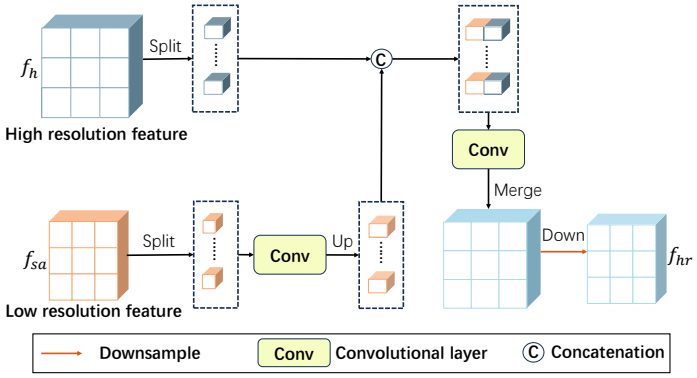

Fig. 3. Schematic diagram of the REG module.

2.2 Shallow Attention Feature Aggregation (SAFA) for accumulated local context enhancement

The transformer model exhibits limitations in capturing the local context. Therefore, we return to the input image to look for additional local features that can complement the features extracted from the transformer encoder. To retain more contextual information about the input image, we first upsample the original image by a scale factor of two. Then the upsampled image is sent to a series of four transposed convolutional layers to produce four feature maps of the same size as the encoded features. The complementary features are concatenated at each level. Inspired by Jun *et al.*[17], which outlines the importance of appearance information in shallow features, we introduce the SAFA module. The combined features are sent to SAFA to gradually enhance the feature's local context. The structure of the SAFA is depicted in Figure 2. It takes four input features f_i,

$i \in 1, 2, 3, 4$, and each f_i has the size of $C_i, \frac{H}{2^{i+1}}, \frac{W}{2^{i+1}}$, where H and W are the height and width of the original input image. All the features first pass a convolutional layer. For easy notation, all convolutional layers are followed by batch normalization and ReLu activation layers which are not depicted separately. After passing the convolutional layers, features are upsampled to the same size as the input image. Then, the features are progressively aggregated from shallow to deep levels, as illustrated in the Figure 2. The combination is accumulative and pairwise, so that after the aggregation three new feature maps are obtained, namely, f'_j for $j \in 1, 2, 3$. Finally, these combined feature maps are concatenated and sent to a convolutional layer to learn the features. The resulting output has the same number of channels as the number of classes to be segmented. To summarise, the output feature map f_{sa} from the SAFA module is formulated as:

$$
\begin{aligned}
f'_1 &= Up(Conv(f_1)) \times Up(Conv(f_2)) \\
f'_2 &= f'_1 \times Up(Conv(f_3)) \\
f'_3 &= f'_2 \times Up(Conv(f_4)) \\
f_{sa} &= Conv([f'_1, f'_2, f'_3])
\end{aligned}
$$

where $Up(\cdot)$ represents the upsampling function, $Conv(\cdot)$ represents the convolutional layer with batch normalization followed by a ReLU activation, and $[\cdot]$ represents concatenation.

2.3 Resolution expansion guidance module

For medical image segmentation, precise edge and boundary delineation are crucial. High-resolution images incorporate finer details about the shape and texture of the target object, which can serve as an excellent resource for refining the coarse segmentation mask. Motivated by the benefits offered by high-resolution images, we design the REG module to facilitate the interaction between high-resolution features and the coarse segmentation mask and thus improve the segmentation mask with finer details along the boundary. In addition, the REG module efficiently utilizes contextual information from the high-resolution image while minimizing GPU memory usage. The module operates through two distinct paths: the low-resolution and high-resolution paths. The high-resolution path processes a large resolution feature, f_h, while the low-resolution path handles the low-resolution feature, f_{sa}, obtained from SAFA. Initially, both features are split into small patches. For the low-resolution patches, the process begins with a pass through a convolutional layer, followed by upsampling to match the size of f_h patches. Subsequently, patches of f_{sa} and f_h are concatenated, and each combined patch passes through another convolutional layer. Finally, the output feature of the REG, f_{hr}, is obtained by merging the patches to restore the original resolution, followed by a downsampling to align with the size of f_{sa}.

2.4 Feature aggregation

In the proposed network, two segmentation maps are obtained, one f_{sa} from the SAFA module and another f_{hr} from the REG module, and aggregated by

addition. Thus, the output segmentation map of SANGRE is represented as:

$$f_{out} = f_{sa} + f_{hr}$$

The final feature combines the benefits of transformer encoding, which extracts global contextual information, with local context features accumulation by the SAFA module. It is further enhanced by detailed visual features acquired through REG. Experiments show that overall feature leads to superior segmentation performance.

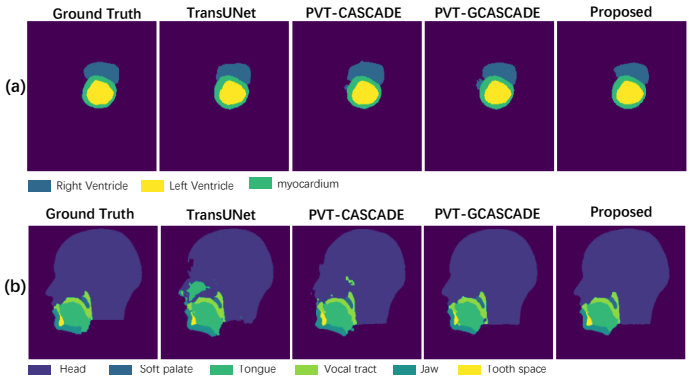

Fig. 4. Comparison of the segmentation results of different models on (a) ACDC dataset and (b) speech MRI dataset.

3 Experiments

3.1 Datasets and implementation

Speech MRI Dataset: contains five series of 105, 71, 71, 78 and 67 images of the upper vocal tract during speech and corresponding segmentation ground truth[18, 19]. Each image contains 6 classes, namely, head, jaw, soft palate, tongue tooth-space and vocal tract. For the experiment on the speech MRI dataset, five-fold cross-validation is used, in which each fold trains on four series and tests on the remaining one, ensuring the subject for test is always different. **Automatic Cardiac Diagnosis Challenge (ACDC) Dataset[20]**: consists of 100 cardiac MRI scans of different patients. Each scan has three classes, which are the right ventricle (RV), spleen (SP) and stomach myocardium (Myo). Following MT-UNet [21], 70 cases (1304 axial slices) are used for training, 10 cases(182 axial slices) are used for validation and 20 cases are used for testing.

3.2 Implementation detail

The implementations of the models were all based on a Nvidia A5000 graphic card and Pytorch 1.10 to allow a consistent comparison of the models' performance. All models were repeated three times and the averaged result is reported. The models were trained for 150 epochs on the speech MRI dataset and 300 epochs on ACDC. All training images have size 256×256. The AdamW optimiser was employed [22]. The hyperparameters for optimization were set to weight decay = 0.00001. A combination of Binary cross entropy (BCE) and Dice loss was used as the loss function. The training and validation data batch sizes were set to be 16 and 1, respectively. The learning rate was set as 0.0003.

Table 1. Comparison of performance of the diffrent models on the speech MRI dataset. DICE Scores (%) in gray and Hausdorff distance (mm).

Model	Head	Jaw	Soft-palate	Tongue	Tooth-space	Vocal-tract	Mean
UNet [1]	99.02	96.86	96.66	98.23	96.03	97.05	97.31
Att-UNet [23]	99.17	96.73	97.35	97.84	97.00	96.74	97.47
TransUNet [11]	99.46	96.65	96.66	98.31	95.53	96.70	97.20
PVT-CASCADE [14]	99.22	97.18	97.54	98.51	95.60	96.32	97.40
PVT-GCASCADE [24]	99.50	97.36	97.56	98.77	95.65	95.65	97.44
Proposed w/o REG	99.47	96.75	96.35	98.29	95.69	96.16	97.12
Proposed w/o SAFA	99.52	97.36	96.83	98.67	96.85	96.59	97.64
Proposed	**99.61**	**97.86**	**97.71**	**98.98**	**97.30**	**97.06**	**98.09**
UNet [1]	18.77	9.36	6.37	12.09	4.38	21.89	12.14
Att-UNet [23]	9.52	8.54	2.29	6.28	2.39	6.82	5.97
TransUNet [11]	16.34	7.66	3.60	18.65	2.75	14.00	10.84
PVT-CASCADE [14]	7.35	1.77	1.15	3.31	1.99	5.18	3.46
PVT-GCASCADE [24]	7.00	7.55	2.46	3.61	3.69	12.59	6.15
Proposed w/o REG	**3.20**	2.10	2.96	3.45	1.72	3.64	2.84
Proposed w/o SAFA	3.66	1.79	1.80	2.76	1.24	3.57	2.47
Proposed	4.05	**1.66**	**1.09**	**2.37**	**1.08**	**3.11**	**2.23**

3.3 Result and ablation study

Figure 4a shows a sample of the qualitative results on the ACDC dataset. The SANGRE network, as proposed, more precisely preserves the shape of the right ventricle compared to alternative methods. For speech MRI, as it is demonstrated by the example in Figure 4b, the segmentation masks generated by the proposed model uphold the structures of various classes with enhanced detail and better precision. Moreover, the occurrence of outlier false classifications is significantly reduced by the proposed method, especially within the head class. Table 1 displays the quantitative performance results of several models on the speech MRI dataset. The proposed SANGRE network outperforms other SOTA approaches across all classes both in terms of the Dice coefficient and Hausdorff distance. Specifically, a mean Dice coefficient of 98.09 % is achieved, which is

roughly 0.6% higher than that of PVT-GCASCADE. Besides, our model exhibits the lowest average Hausdorff distance of 2.23 compared to other methods, demonstrating its capability to capture fine structural details. On the ACDC dataset, similarly, a clearly advantage in Dice coefficient is observed comparing against the SOTA PVT-GCASCADE, with a comparable number of parameters and FLOPS, as presented in Table 2. The performance of the proposed network excluding the REG and SAFA modules is also presented in Tables 1, and 2. It is demonstrated in Table 1 that the employment of REG in the model significantly reduces the Hausdorff distance in small feature classes, such as the jaw, soft palate, and tooth space. Overall, the superior performance of the full model on both datasets underscores the effectiveness of both proposed modules.

Table 2. Comparison of the performance of the different models on the ACDC dataset.

Model	Avg Dice (%)	RV	Myo	LV	#Params (M)	FLOPs (G)
R50+UNet [11]	87.55	87.10	80.63	94.92	-	-
ViT+CU[11]	81.45	81.46	70.71	92.18	-	-
TransUNet [11]	89.71	86.67	87.27	95.18	105.28	24.64
MT-UNet [21]	90.43	86.64	89.04	95.62	-	-
MISSFormer [10]	90.86	89.55	88.04	94.99	-	-
PVT-CASCADE [14]	91.46	89.57	88.9	94.50	34.13	5.84
PVT-GCASCADE [24]	91.95	90.31	89.63	95.91	26.64	**4.252**
Proposed w/o REG	91.45	89.28	89.41	95.66	**24.89**	5.29
Proposed w/o SAFA	91.57	89.38	89.61	95.73	25.87	5.34
Proposed	**92.29**	**90.92**	**89.94**	**96.02**	25.87	5.48

4 Conclusion

In this paper, we present a novel network SANGRE for MR image segmentation. The network features two key modules: the Shallow Attention Feature Aggregation module progressively eliminates irrelevant noise from the background of the features, enriching the local context in the process; while the resolution expansion guidance module enhances the feature map by leveraging the feature's appearance information from images with expanded resolution, ensuring the preservation of small-sized feature classes. While resolution expansion is involved, the REG module only increases the FLOPs by 0.19G. Our experimental results, conducted on the speech MRI and ACDC datasets, demonstrate that our network surpasses other SOTA methods in all aspects.

Acknowledgments. Ying He is funded by Barts Charity under Grant G-002066.

Disclosure of Interests. None of the authors have any conflicts of interest to declare.

References

1. Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.
2. 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.
3. Tiejun Yang, Yudan Zhou, Lei Li, and Chunhua Zhu. Dcu-net: Multi-scale u-net for brain tumor segmentation. *Journal of X-ray Science and Technology*, 28(4):709–726, 2020.
4. Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE transactions on medical imaging*, 39(6):1856–1867, 2019.
5. Huimin Huang, Lanfen Lin, Ruofeng Tong, Hongjie Hu, Qiaowei Zhang, Yutaro Iwamoto, Xianhua Han, Yen-Wei Chen, and Jian Wu. Unet 3+: A full-scale connected unet for medical image segmentation. In *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 1055–1059. IEEE, 2020.
6. 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.
7. 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.
8. Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 568–578, 2021.
9. Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision*, pages 205–218. Springer, 2022.
10. Xiaohong Huang, Zhifang Deng, Dandan Li, Xueguang Yuan, and Ying Fu. Missformer: An effective transformer for 2d medical image segmentation. *IEEE Transactions on Medical Imaging*, 2022.
11. 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.
12. Feiniu Yuan, Zhengxiao Zhang, and Zhijun Fang. An effective cnn and transformer complementary network for medical image segmentation. *Pattern Recognition*, 136:109228, 2023.
13. 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: 24th International Conference, Stras-*

bourg, France, September 27–October 1, 2021, Proceedings, Part I 24, pages 14–24. Springer, 2021.
14. Md Mostafijur Rahman and Radu Marculescu. Medical image segmentation via cascaded attention decoding. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6222–6231, 2023.
15. Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvt v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 8(3):415–424, 2022.
16. Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
17. Jun Wei, Yiwen Hu, Ruimao Zhang, Zhen Li, S Kevin Zhou, and Shuguang Cui. Shallow attention network for polyp segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24*, pages 699–708. Springer, 2021.
18. M. Ruthven, A. Peplinski, D.M. Adam, A.P. King, and M.E. Miquel. Real-time speech mri datasets with corresponding articulator ground-truth segmentations.[data descriptor]. *Scientific Data*, 10:10.1038/s41597–023–02766–z, 2023.
19. M. Ruthven, A. Peplinski, and M. Miquel. A multi-speaker dataset of real-time two- dimensional speech magnetic resonance images with articulator ground-truth segmentations, [dataset]. *Zenodo*, page 10.5281/zenodo.10046815, 2023.
20. Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, et al. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE transactions on medical imaging*, 37(11):2514–2525, 2018.
21. 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.
22. Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.
23. Jo Schlemper, Ozan Oktay, Michiel Schaap, Mattias Heinrich, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention gated networks: Learning to leverage salient regions in medical images. *Medical image analysis*, 53:197–207, 2019.
24. Md Mostafijur Rahman and Radu Marculescu. G-cascade: Efficient cascaded graph convolutional decoding for 2d medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 7728–7737, 2024.