Max-Fusion U-Net for Multi-Modal Pathology Segmentation with Attention and Dynamic Resampling

Citation for published version:
Jiang, H, Wang, C, Chartsias, A & Tsaftaris, S 2020, Max-Fusion U-Net for Multi-Modal Pathology Segmentation with Attention and Dynamic Resampling. in *MyoPS: Myocardial Pathology Segmentation Combining Multi-Sequence CMR Challenge: Myocardial Pathology Segmentation Combining Multi-Sequence Cardiac Magnetic Resonance Images.* Lecture Notes in Computer Science , vol. 12554, Springer, pp. 68-81, MICCAI 2020 , 4/09/20. https://doi.org/10.1007/978-3-030-65651-5_7

Digital Object Identifier (DOI):
10.1007/978-3-030-65651-5_7

Link:
Link to publication record in Edinburgh Research Explorer

Published In:
MyoPS: Myocardial Pathology Segmentation Combining Multi-Sequence CMR Challenge

Max-Fusion U-Net for Multi-Modal Pathology Segmentation with Attention and Dynamic Resampling

Haochuan Jiang[1], Chengjia Wang[2]✉, Agisilaos Chartsias[1], and Sotirios A. Tsaftaris[1,3]

[1] School of Engineering, University of Edinburgh, U.K.
[2] Center for Cardiovascular Science, University of Edinburgh, U.K.,
chengjia.wang@ed.ac.uk
[3] The Alan Turing Institute, U.K.

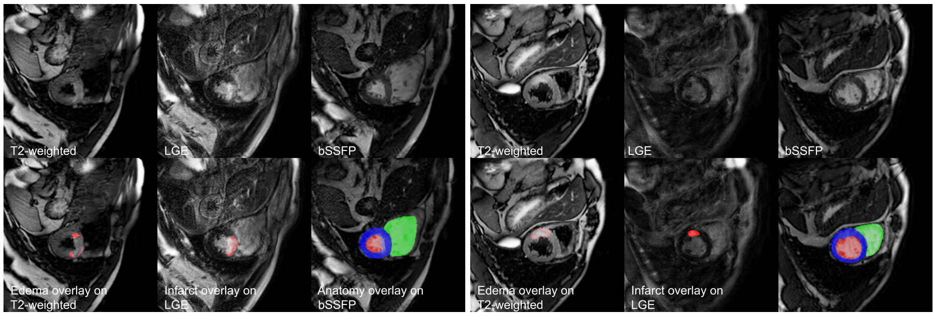
Abstract. Automatic segmentation of multi-sequence (multi-modal) cardiac MR (CMR) images plays a significant role in diagnosis and management for a variety of cardiac diseases. However, the performance of relevant algorithms is significantly affected by the proper fusion of the multi-modal information. Furthermore, particular diseases, such as myocardial infarction, display irregular shapes on images and occupy small regions at random locations. These facts make pathology segmentation of multi-modal CMR images a challenging task. In this paper, we present the Max-Fusion U-Net that achieves improved pathology segmentation performance given aligned multi-modal images of LGE, T2-weighted, and bSSFP modalities. Specifically, modality-specific features are extracted by dedicated encoders. Then they are fused with the pixel-wise maximum operator. Together with the corresponding encoding features, these representations are propagated to decoding layers with U-Net skip-connections. Furthermore, a spatial-attention module is applied in the last decoding layer to encourage the network to focus on those small semantically meaningful pathological regions that trigger relatively high responses by the network neurons. We also use a simple image patch extraction strategy to dynamically resample training examples with varying spacial and batch sizes. With limited GPU memory, this strategy reduces the imbalance of classes and forces the model to focus on regions around the interested pathology. It further improves segmentation accuracy and reduces the mis-classification of pathology. We evaluate our methods using the Myocardial pathology segmentation (MyoPS) combining the multi-sequence CMR dataset which involves three modalities. Extensive experiments demonstrate the effectiveness of the proposed model which outperforms the related baselines.

Keywords: pathology segmentation · multi-modal · max-fusion · dynamic resample

1 Introduction

Cardiac diseases are typically assessed using multiple cardiac MR (CMR) sequences (modalities), providing complementary information. For example, Late Gadolinium Enhancement (LGE) detects myocardial infarct, T2-weighted (T2) images provide clear visibility of acute injury and ischemic regions, and balanced-Steady State Free Precession cine sequence (bSSFP) offers high contrast between anatomical regions and captures cardiac motion.

Deep learning models have been extensively used for automatic segmentation of multi-modal data. A critical step for the analysis of multi-modal CMR data is to effectively fuse information from multiple modalities. Prior works [5] concatenate the feature maps extracted from different modalities into different channels and fuse them in the following convolutional layers. Other methods [9, 3] merge the features across different layers of the neural network, where a cross-modal convolution fusion model is introduced in [13]. In [7] they employ dedicated encoders for different modalities to encode different types of information, for example, content and style features from the corresponding input data. The features are then fused using channel concatenation in the U-Net skipping-connections. A similar idea was also used in [1] where a maximum fusion operator instead of simple concatenation in the skip-connections is applied on disentangled anatomy factors extracted from different modalities at the end of encoders.



(a) Example 1 with anatomy overlay (b) Example 2 with anatomy overlay

Fig. 1: Examples of multi-modal CMR images overlaying anatomy and pathology.

One other challenge in segmenting pathology such as myocardial infarct and edema is that these pathologies are often of diverse shape and occur at random positions. As such, shape priors such as mask discriminator [1] cannot be used. Besides, the interested pathology and anatomy only occur within a small region of the whole image, as examples of multi-sequence CMR images with manually segmented anatomy and pathology (myocardial infarction and edema) given in Fig. 1. This makes the data distribution highly imbalanced across classes, resulting in overfitting in the training data. Particularly in current popular backbone convolutional neural networks (CNN) assuming all pixels in the image contribute equally to the final prediction, the over-fitting issue is even worse. A possible solution is to use the spatial attention module [4], leading the network to focus on specific image regions. In our case, the focus corresponds to pathology pixels.

In addition, given limitations in GPU memory, training can only be performed with a small batch size. This even worsen the overfitting issue since due to this and small pathological region in each image, in each training iteration, only a small amount of pathology pixels are seen by the network. Nevertheless, the batch size can be increased if training with smaller size patches instead of full images, e.g., by engaging random cropping. Although it is commonly used as a data augmentation technique [12], all patches are treated equally importantly. It is appealing if the cropping strategy will oversample patches around pathology regions that we are interested in.

In this paper, we propose the Max-Fusion U-Net (MFU-Net) for cardiac pathology segmentation, given fully-annotated multi-modal aligned images. We use dedicated encoders to extract features for each modality, as in [1, 7]. But rather than channel concatenation [7], we fuse features from different modalities with the pixel-wise maximum operator applied on each layer [1]. This fusion operator guides the network to keep informative features extracted by each modality. At the same time, fusion with maximum operator indirectly encourages feature maps to encode important features in high intensities including pathological pixels. A spatial-attention module is also employed in the last decoding layer to modulate the spatial focus, which in our case means to increase focus of the pathology pixels. Finally, to address the issue that only a small amount of pathological pixels are exposed to the network during training, we adopt a dynamic resampling strategy. To obtain each batch, we extract multiple patches around the interested pathology based on an arbitrary probability, then extract the rest data by randomly cropping the image to the same size. By feeding more patches related to pathological regions and less related to background patches, the network will thus naturally become more sensitive to pathological pixels. At the same time, the training batch size can be dynamically enlarged without occupying extra computation resources due to the reduced image dimension. Theoretically, the spatial size of the training data should not harm the training efficiency as long as the sampled image patches are bigger than the largest receptive field of the network. Extensive experiments have demonstrated the effectiveness of the proposed MFU-Net in cardiac pathology segmentation including infarction and edema when given multi-modal inputs including LGE, T2-weighted, and bSSFP, outperforming relevant methods. Major **contributions** of this work are summarized as follows:

- We proposed the MFU-Net that fuses multi-modal features extracted by dedicated encoders with the pixel-wise maximum operator;
- We incorporate a spatial-attention module to guide the network to focus on the pathology region;

- We proposed a novel training strategy by feeding randomly resampled subpatches from the original training data with more probability around the pathology region, at the same time increasing the batch size dynamically;
- MFU-Net improves the Dice score of state-of-the-art benchmarks on myocardial pathology segmentation on multi-modal CMR 2020 dataset [15, 16].

2 Methodology

This section presents the proposed MFU-Net model, and the details about the architecture, the modality-specific encoders, the maximum fusion operator, and the attention-based decoding modules.

Overview: Let $X_{LGE}, X_{T2}, X_{bSSFP}$ represent images of LGE, T2-weighted, and bSSFP CMR modalities respectively, and Y_{ana}, Y_{pat} be the associated anatomy and pathology masks. If i enumerates all samples from the above sets, we assume a fully labelled multi-modal pathology subset $\mathbf{L} = \{x^i_{LGE}, x^i_{T2}, x^i_{bSSFP}, y^i_{ana}, y^i_{pat}\}$, where three modality slices $x^i_{LGE}, x^i_{T2}, x^i_{bSSFP} \subset \mathbb{R}^{H\times W}$ are preprocessed [15, 16], such that they are aligned in a common space and are resampled to the same spatial resolution. In addition, $y^i_{ana} \in Y_{ana} := \{0,1\}^{H\times W\times N}$, and $y^i_{pat} \in Y_{pat} := \{0,1\}^{H\times W\times K}$, where N and K denote the number of anatomy, and pathology masks respectively.[4], and H and W are the image height and width.

2.1 Model Architecture

The architecture of MFU-Net is illustrated in Fig. 2. It consists of three modality-specific encoders, a multi-modal feature fusion with pixel-wise maximum operator, and a decoder with a spatial attention module that produces the segmentation results.

Individual Encoders: The original U-Net architecture [10] only specifies a single encoder to extract features. To accommodate differences in the pixel intensity distributions between modalities, we expand the U-Net by using one independent encoder for each modality. This leads to three modality-specific encoders. Represented by red, green, and blue colors in Fig. 2, these encoders are denoted as Enc_{LGE}, Enc_{T2}, and Enc_{bSSFP} respectively for LGE, T2, and bSSFP data. The encoded features $Enc_{LGE}(x^i_{LGE})$, $Enc_{T2}(x^i_{T2})$, and $Enc_{bSSFP}(x^i_{bSSFP})$ are concatenated and used as input to the bottleneck blocks (the transparent brown blocks in Fig. 2).

Modality Fusion: A simple way for feature fusion is through channel concatenation [7]. However, this strategy does not really merge the modality-specific information into modality-independent features, so that the contribution of different modalities can not be balanced dynamically. Such adaptive balancing among modalities is particularly important in pathology segmentation, where specific pathologies can only be spot in particular modalities, i.e. infarct can only be seen in LGE, while edema can only be seen in T2, as seen in Fig. 1.

[4] We restrict to the case where $N = 3$ (myocardium, left ventricle, and right ventricle) and $K = 2$ (infarction and edema).

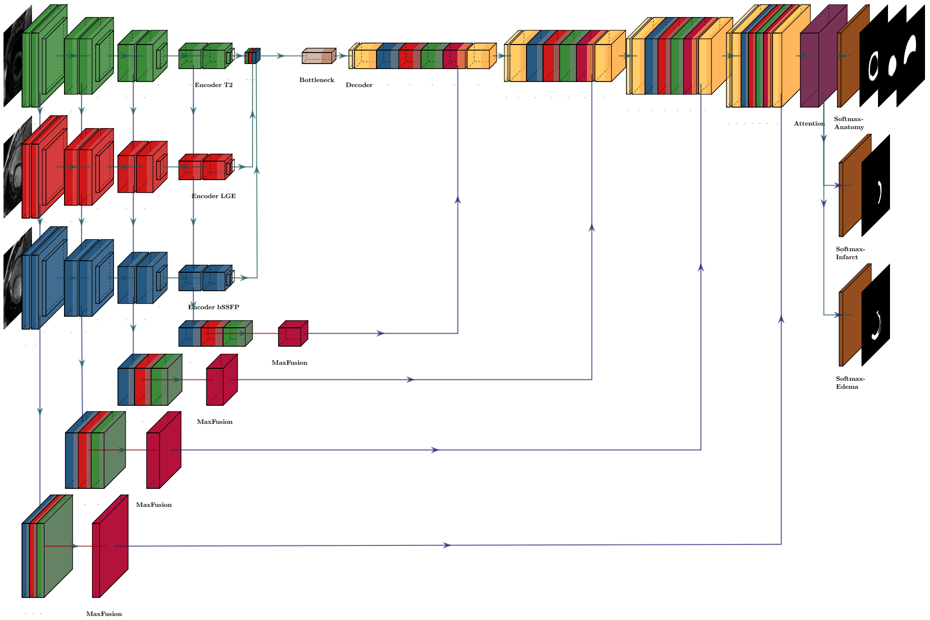


Fig. 2: MFU-Net Architecture. Red, Green, and Blue blocks represent LGE, bSSFP, and T2 encoding features. Yellow blocks depict the decoding features. Pink Blocks are max-fused features, while transparent brown block is the bottleneck feature. Solid brown ones are the softmaxed probability map, while the amaranth block is the spatial attention module.

Instead, we would like to fuse the feature in an auto-selective fashion. To this end, we employ the pixel-wise maximum operator, which has been previously used in [1] for dual-modal anatomy segmentation. In the proposed MFU-Net, the fusion is among features generated by the dedicated encoders, producing the fused feature as depicted in pink blocks in Fig. 2. Rather than fusing latent features of one layer [1], we apply the max-fusion operation to different blocks in the encoders for multi-scale mixture of the multi-modal information. For instance, for the k-th encoding layer, the fusion is performed by $Enc^k(x^i_{LGE}, x^i_{T2}, x^i_{bSSFP}) = \max(Enc^k_{LGE}(x^i_{LGE}), Enc^k_{T2}(x^i_{T2}), Enc^k_{bSSFP}(x^i_{bSSFP}))$ in a pixel-wise fashion.[5] It provides the dynamically selective features across modalities. However, the conventional concatenation features do not differentiate features from different modalities. The fused feature Enc^k, together with the linear concatenation of $Enc^k_{LGE}(x^i_{LGE})$, $Enc^k_{T2}(x^i_{T2}$, and $Enc^k_{bSSFP}(x^i_{bSSFP})$, are then concatenated to the corresponding decoding layer with a skip connection, as in the original U-Net [10]. The linear concatenated and nonlinear max-fused representations provide the complementary information for the modal-specific features.

[5] For simplicity we note it as Enc^k in following sections.

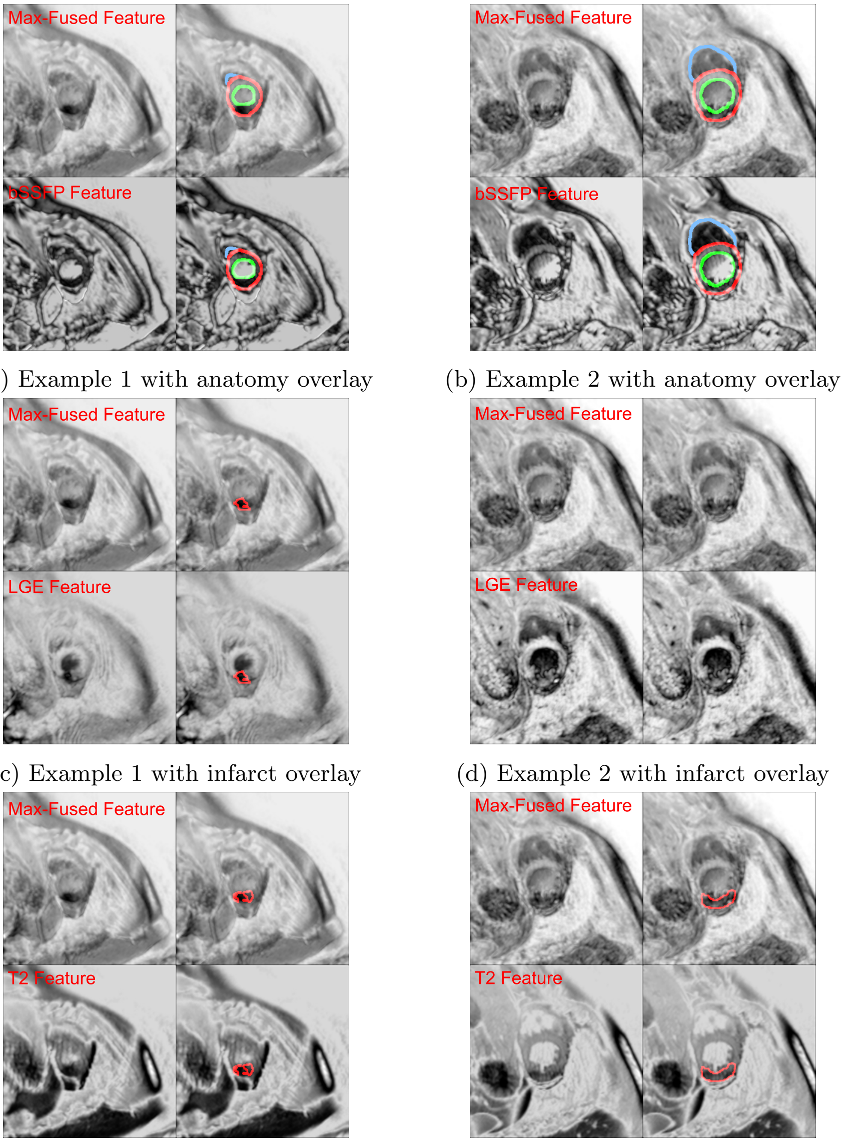
Two examples of the max-fused features compared to single-modal features are shown in Fig. 3. As discussed above, specific pathologies can only be observed in particular modalities clearly. For example, myocardial infarction can only be observed on features extracted from LGE data as a small dark area (Fig. 3c and 3d). Similarly, the boundary of edema can only be depicted on T2 feature maps (Fig. 3e and 3f). In comparison, both pathological regions can be easily detected with relatively clearer boundaries on the max-fused feature maps (Fig. 3c, 3d, 3e, 3f). Furthermore, the interested anatomical structures can be seen as easily as in bSSFP features (Fig. 3a and 3b). On the contrary, the boundaries of heart anatomy and edema are blurred in LGE, so is the infarction in T2 data. Boundaries of both infarction and edema are hard to be detected in bSSFP data. This can be seen as a qualitative evidence of an effective mixture of the multi-modality information.

Decoding with Attention: The decoder of MFU-Net receives as input the bottleneck layer that follows the concatenated multi-modal features of the encoding part. A series of convolutional blocks upsample the spatial resolution as in U-Net, and are concatenated with the encoding features (including the max-fused feature and the corresponding encoding features for each modality) computed at the corresponding layers of the encoder with skip connections.

Since cardiac pathologies often occupy in a small part of the whole image, producing segmentations by treating each pixel equally is challenging and might lead the network to concentrate more on the background but ignore tiny pathological regions. In order to overcome this issue, we use a spatial attention mechanism [4] to capture long-range pixel dependencies and assign different weights on different regions. In this sense, segmentation can be improved by selecting useful information in features extracted around the pathological regions and by discarding unrelated features. In detail, the spatial attention module, shown in Fig. 4, is applied at the last layer of the decoding path with the architecture of the spatial and channel attention modules following [4]. In order to reduce computational complexity introduced when the feature dimensions are large, we first downsample the input feature using stride-2 convolutions before calculating the query, key, and value tensors. The attention module is depicted in Fig. 4. After calculating the attention map, the dimension will be recovered by deconvolution in the upsampling block. Fig. 5 gives examples of spatial attention outputs with corresponding predicted masks. Clearly, the corresponding mask region is highlighted in the spatial attention maps, demonstrating the utility of this mechanism in segmentation.

3 Implementation

In this section, the implementation details of the proposed MFU-Net will be specified. Firstly, we will introduce the dynamic resampling training strategy, then the alternative cross-validation to make full use of the training data and avoid overfitting issue will be specified.



(a) Example 1 with anatomy overlay

(b) Example 2 with anatomy overlay

(c) Example 1 with infarct overlay

(d) Example 2 with infarct overlay

(e) Example 1 with edema overlay

(f) Example 2 with edema overlay

Fig. 3: Two examples of comparison between the feature maps extracted before and after the max-fusion operation in terms of visibility of: (a) and (b) anatomy; (c) and (d), myocardial infarction; (e) and (f) edema. For each subfigure, the max-fused feature maps are shown at the top and modality-specific feature maps are shown at the bottom. The object boundaries overlapped with the feature maps are on the right.

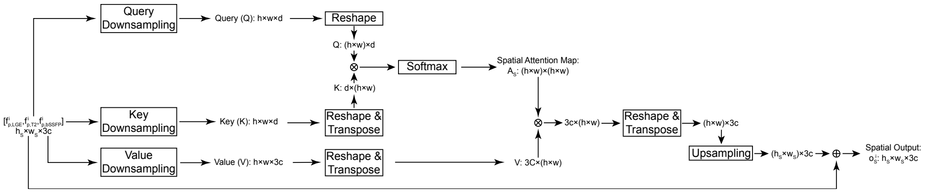


Fig. 4: Attention Module at the last decoding layer. $\oplus$ and $\otimes$ represent element-wise summation and multiplication respectively between two matrices.

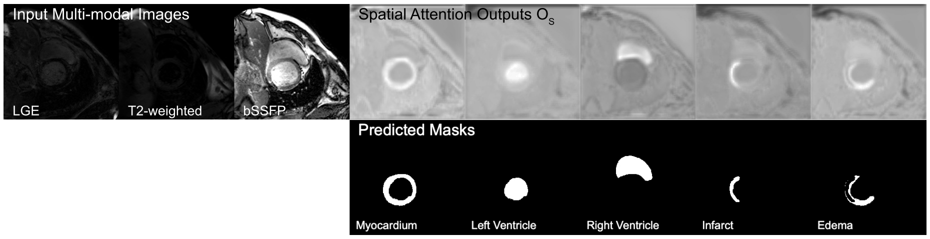


Fig. 5: Spatial attention outputs correspond to the predicted masks.

3.1 Dynamic Resampling Training Strategy

The proposed MFU-Net is deployed on a GTX Titan X GPU with 12GB standard memory. In the training process, the available memory allows 288×288 image size with a batch size equals to 4. In order to increase the model's focus on pathological regions, we also train with patches of different sizes that are dynamically resampled. For the batch obtained at the t-th iteration, we first decide the patch size d_t by $d_t = 96 + 16i$, $i \in \{1, \cdots, 12\}$ where i is randomly picked. Then, with an arbitrary probability ρ_c, an extracted patch is centred on the pathology of interest. The dynamic batch size N_t is decided by $N_t = \lfloor d_{t-1}^2 N_{t-1} / d_t^2 \rfloor$. For example, in the first iteration, we initialize the image size $d_0 = 288$, thus when extracting 96×96 image patches, the batch size can be as big as 36. This not only increases the batch size but also allows to manual balance the data distribution. In this work, we set $\rho_c = 0.89$ as the interested anatomy only takes up 11% pixels of the whole image. As such, pathological regions are more probably to be seen in the cropped patches. Fig. 6 demonstrates the details of this sampling process with two different patch sizes.

3.2 Training with Alternative Cross Validation

To make full use of the training data and avoid possible overfitting issues, we employ an alternative cross-validation strategy as part of training to predict the MyoPS 2020 challenge testing data. Specifically, the whole training set is split into five parts. Accordingly, the training process will be specified in five phases. In each phase, four out of the five splits are selected as the training set, while

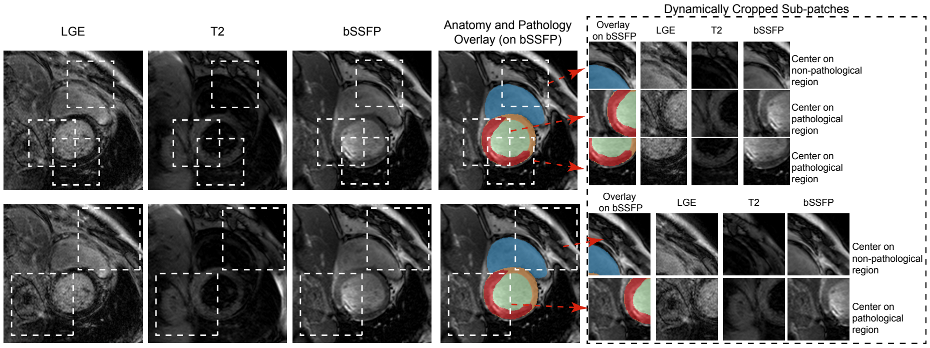

LGE
T2
bSSFP
Anatomy and Pathology Overlay (on bSSFP)
Dynamically Cropped Sub-patches
Overlay on bSSFP
LGE
T2
bSSFP
Center on non-pathological region
Center on pathological region
Center on pathological region
Overlay on bSSFP
LGE
T2
bSSFP
Center on non-pathological region
Center on pathological region

Fig. 6: Dynamic resampled image patches with varying spatial and batch sizes. The resampling sizes for images in the first and second row is 96 and 128 respectively. Smaller resampling size will bring greater batch size.

the remaining one is used as the validation set to prevent overfitting by defining the early-stopping criteria. If one phase of training is terminated, the network optimization continues on another split. The number of epochs for each training phase are 50, 40, 30, 20, and 15, while the initial learning rate are 0.0001, 0.00009, 0.00008, 0.00006, 0.00005 respectively and decayed exponentially. When all the five training phases are completed, we add a final fine-tuning phase that involves all the training data but is trained only 10 epochs with the small learning rate at 0.00004 and decayed exponentially as well. This will avoid the model to forget early trained examples.

4 Experiments

We evaluate the proposed MFU-Net on pathology segmentation using the Dice score. Experimental setup, datasets, benchmarks, and training details will be detailed in the following part.

Data: We evaluate our proposed MFU-Net on the multi-sequence CMR (MyoPS) dataset [15, 16] that contains in total 25 volumes and 102 slices in the training set. For each slice, three modalities including LGE, T2, and bSSFP are provided. They are preprocessed with the Multi-variate Mixture Model [15, 16], such images from the three modalities are aligned and resampled to same spatial resolution. For all the images, three anatomy masks (myocardium, left ventricle, and right ventricle) and two pathology masks (myocardial infarct and edema) are given. The testing set contains 20 volumes and 72 slices without ground-truth masks available. Both training and testing data are cropped to 288 × 288 to keep the region to be segmented in the sight.

Training details: The proposed MFU-Net is optimized with fully supervised losses. The segmentation of both anatomy and pathology is trained with tversky [11] and focal [8] losses in a supervised fashion. The tversky loss is defined as $\ell_{T,j} = (\hat{y}^i_j \odot y^i_j)/[\hat{y}^i_j + y^i_j + (1-\beta)\cdot(\hat{y}^i_j - \hat{y}^i_j \odot y^i_j) + \beta\cdot(y^i_j - \hat{y}^i_j \odot y^i_j)]$ and the focal loss

Table 1: Anatomy and pathology segmentation dice scores (%) of MFU-Net and relevant variants with *Residual* backbone. Myo., LV, and RV represent the myocardium, left ventricle, and right ventricle respectively. *max*, *attention*, and *resample* represent the presence of the max-fusion operator, the spatial attention module, and the dynamic resampling strategy respectively. Pathology score is calculated by averaging both the infarct and edema segmentation performance.

max	*attention*	*resample*	Myo.	LV	RV	Infarct	Edema	Avg. Pathology
✓	✓	✓	$84.3_{7.9}$	$\mathbf{87.5_{7.1}}$	$78.5_{14.2}$	$\mathbf{53.0_{20.5}}$	$28.7_{13.9}$	$\mathbf{44.9_{13.9}}$
–	✓	✓	$\mathbf{85.2_{8.1}}$	$86.8_{10.6}$	$\mathbf{78.7_{14.0}}$	$52.1_{20.4}$	$\mathbf{29.4_{12.4}}$	$42.9_{14.0}$
✓	–	✓	$84.2_{6.9}$	$86.9_{7.5}$	$76.7_{13.7}$	$46.1_{21.1}$	$28.1_{14.3}$	$41.0_{14.1}$
✓	✓	–	$84.5_{5.3}$	$87.1_{6.4}$	$74.9_{18.7}$	$49.4_{20.4}$	$\mathbf{29.4_{17.9}}$	$42.8_{15.5}$
–	–	✓	$81.1_{7.8}$	$84.2_{8.0}$	$67.2_{17.2}$	$50.2_{17.8}$	$19.3_{13.0}$	$37.5_{16.8}$
–	✓	–	$\mathbf{85.2_{4.1}}$	$86.1_{9.4}$	$75.7_{18.6}$	$52.6_{19.4}$	$28.7_{17.1}$	$43.6_{15.4}$
✓	–	–	$82.3_{7.9}$	$82.3_{8.9}$	$68.2_{16.5}$	$48.0_{25.4}$	$22.8_{15.9}$	$36.1_{18.5}$
–	–	–	$81.6_{6.5}$	$84.1_{8.0}$	$67.5_{15.5}$	$42.8_{21.7}$	$20.6_{16.6}$	$34.8_{17.7}$

is $\ell_{F,j} = \sum_{H,W}[-y_j^i(1-\hat{y}_j^i)^\gamma log(\hat{y}_j^i)]$, where $\odot$ represents the element-wise multiplication and j corresponds to the involved anatomy or pathology labels. We set penalties for anatomy, infarct, and edema equal to $\lambda_{anatomy} = 1$, $\lambda_{infarct} = 3$, and $\lambda_{ana} = 5$ respectively, for each of the tversky and focal losses. Moreover, in order to achieve more stable training and quicker convergence, we initialise MFU-Net with weights from the MMSDNet [1] encoder (that also follows a U-Net architecture with dedicated encoders for each modality) when trained only with the unsupervised reconstruction loss.

Benchmarks: We evaluate the pathology segmentation performance of MFU-Net using several variants of our model. More specifically, we evaluate the effect of different design choices including the maximum fusion operator, the spatial attention module and the dynamic resampling strategy. In total we construct eight ablated models, all of which concatenate features at each encoding layer.

4.1 Results and Discussion

We report segmentation results of MFU-Net and the ablated models in Table 1 with anatomy (myocardium, left and right ventricles) and pathology (myocardial infarct and edema) segmentation dice scores.[6] The backbone architecture used the residual connections in encoding and decoding layers [6] noted as *Residual*.

As can be seen in Table 1, the proposed maximum fusion operator and dynamic resampling achieve the best infarct segmentation, while edema segmentation performs similarly to the model without the max fusion. On average the

[6] Since we do not have the ground truth of the testing data, the performance reported in Table 1 and Table 2 are obtained by five-fold cross validation across the training set. Relevant splits are following the description in Sec. 3.2. In addition, we also report the averaged pathology Dice scores of the both pathologies to assess the overall pathology segmentation performance.

Table 2: Anatomy and pathology segmentation comparison between *Residual*, *Dilation*, and *Sideconv* backbones when *max*, *attention*, and *resample* are all present.

	Myo.	LV	RV	Infarct	Edema	Avg. Pathology
Residual	$\mathbf{84.3_{7.9}}$	$\mathbf{87.5_{7.1}}$	$\mathbf{78.5_{14.2}}$	$53.0_{20.5}$	$28.7_{13.9}$	$\mathbf{44.9_{13.9}}$
Dilation	$80.5_{4.3}$	$85.3_{6.3}$	$44.3_{33.8}$	$\mathbf{55.1_{18.7}}$	$23.1_{13.9}$	$43.7_{14.0}$
Sideconv	$76.3_{10.3}$	$65.0_{18.6}$	$40.5_{39.4}$	$52.1_{21.1}$	$\mathbf{29.7_{11.8}}$	$45.0_{16.0}$

Table 3: Pathology segmentation dice scores on the MyoPS 2020 testing data

SideConv		Dilation	
Infarct	Infarct+Edema	Infarct	Infarct+Edema
$57.0_{28.7}$	$60.3_{18.1}$	$58.4_{26.3}$	$61.4_{17.8}$

model with all *attention*, *max*, and *resample* options achieves the best pathology segmentation with Dice equal to 44.9%.[7] Moreover, it can be observed that the spatial attention module improves segmentation for both infarct and edema.

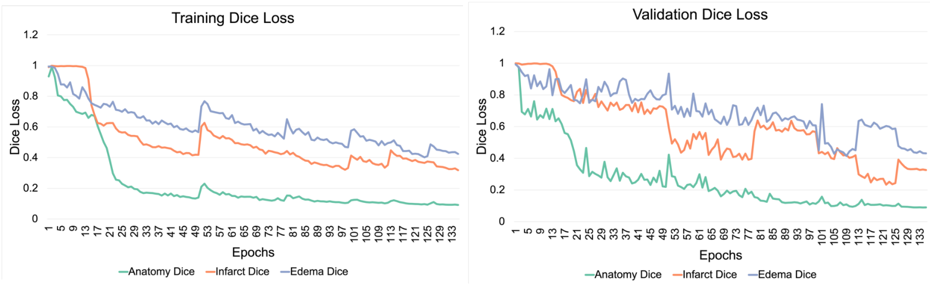
In addition, the anatomy segmentation does not benefit from the proposed compositions, particularly in ventricles. The reason is two-folded. On one hand, the MyoPS 2020 challenge concentrates mainly on the pathology segmentation. As such, during training, we put more penalties on the pathology supervision (Sec. 4). It results in less focus on anatomy learning. On the other hand, because both infarct and edema is in the myocardium region, the pathology training gradient will offer an additional guide to train myocardium segmentation. On the contrary, ventricle predictions are not enjoying such an advantage.

4.2 Prediction for the Challenge Testing Dataset

Table 2 specifies the comparison with other two backbone CNN options, namely, the dilated convolutions in the bottleneck layer [14], and the side-convolution by adding 3×3, 3×1, and 1×3 convolutions in each of the convolution operations [2]. They are denoted as *Dilation* and *SideConv* respectively. It can be seen clearly that the models using dilated convolutions and side-convolutions improve on the segmentation of infarct and edema respectively, compared to our initial model using residual connections. We therefore use the *Dilation* and *Sideconv* MFU-Nets for inference of the MyoPS 2020 testing dataset. The segmentation results are presented in Table 3 and contain the Dice scores of infarct and the union of both infarct and edema. It can be seen that the *dilation* backbone with *max*, *attention*, and *resample* achieves better results with 58.4% dice for infarct, and 61.4% for both the infarct and the edema together.

[7] Although the anatomy segmentation performance decreases, we still think *SideConv* and *Dilation* are better choices since we are more caring about the pathology prediction in this research.

The prediction models are trained with the alternative cross validation described in Sec. 3.2. Fig. 7a and Fig. 7b illustrate the training and validation dice losses respectively during model optimization. Particularly, in Fig. 7a, each loss jump corresponds to the point where the cross validation split switches and the training phase changes. Furthermore, all losses gradually decrease in each training phase, and finally converge at the final few steps.



(a) Training dice losses

(b) Validation dice losses

Fig. 7: Training and validation dice losses with the alternative cross-validation for the testing dataset. Curves in green, orange, and blue represent the anatomy, infarct, and edema dice losses.

5 Conclusions

In this paper, we proposed the Multi-Fusion U-Net, a novel architecture to segment infarct and edema from multi-modal images including LGE, T2-weighted, and bSSFP sequences. Our model uses dedicated encoders for each modality, and combines multi-modal information with feature fusion performed with the pixel-wise maximum operator at each encoding layer. These max-fused features together with the concatenated modality-specific features of each encoding layer, are propagated to corresponding decoding layers of the same spatial resolution using skip connections. Additionally, a spatial attention module in the final decoding layer, as well as a novel dynamic resampling training strategy, are engaged to guide the network to focus on small pathology regions. Extensive experiments on the MyoPS 2020 challenge dataset demonstrated the effectiveness of the MFU-Net in improving cardiac pathology segmentation performance.

Acknowledgement: This work was supported by US National Institutes of Health (1R01HL136578-01). This work used resources provided by the Edinburgh Compute and Data Facility (http://www.ecdf.ed.ac.uk/). S.A. Tsaftaris acknowledges the Royal Academy of Engineering and the Research Chairs and Senior Research Fellowships scheme.

References

1. Chartsias, A., Papanastasiou, G., Wang, C., Semple, S., Newby, D., Dharmakumar, R., Tsaftaris, S.A.: Disentangle, align and fuse for multimodal and zero-shot image segmentation. arXiv preprint arXiv:1911.04417 (2019)
2. Ding, X., Guo, Y., Ding, G., Han, J.: Acnet: Strengthening the kernel skeletons for powerful cnn via asymmetric convolution blocks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1911–1920 (2019)
3. Dolz, J., Gopinath, K., Yuan, J., Lombaert, H., Desrosiers, C., Ayed, I.B.: Hyperdense-net: a hyper-densely connected cnn for multi-modal image segmentation. IEEE transactions on medical imaging **38**(5), 1116–1126 (2018)
4. Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H.: Dual attention network for scene segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3146–3154 (2019)
5. Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., Pal, C., Jodoin, P.M., Larochelle, H.: Brain tumor segmentation with deep neural networks. Medical image analysis **35**, 18–31 (2017)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
7. Jiang, H., Yang, G., Huang, K., Zhang, R.: W-net: one-shot arbitrary-style chinese character generation with deep neural networks. In: International Conference on Neural Information Processing. pp. 483–493. Springer (2018)
8. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017)
9. Mahmood, F., Yang, Z., Ashley, T., Durr, N.J.: Multimodal densenet. arXiv preprint arXiv:1811.07407 (2018)
10. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
11. Salehi, S.S.M., Erdogmus, D., Gholipour, A.: Tversky loss function for image segmentation using 3d fully convolutional deep networks. In: International Workshop on Machine Learning in Medical Imaging. pp. 379–387. Springer (2017)
12. Takahashi, R., Matsubara, T., Uehara, K.: Ricap: Random image cropping and patching data augmentation for deep cnns. In: Asian Conference on Machine Learning. pp. 786–798 (2018)
13. Tseng, K.L., Lin, Y.L., Hsu, W., Huang, C.Y.: Joint sequence learning and cross-modality convolution for 3d biomedical segmentation. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 6393–6400 (2017)
14. Vesal, S., Ravikumar, N., Maier, A.: A 2d dilated residual u-net for multi-organ segmentation in thoracic ct. arXiv preprint arXiv:1905.07710 (2019)
15. Zhuang, X.: Multivariate mixture model for cardiac segmentation from multi-sequence mri. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 581–588. Springer (2016)
16. Zhuang, X.: Multivariate mixture model for myocardial segmentation combining multi-source images. IEEE transactions on pattern analysis and machine intelligence **41**(12), 2933–2946 (2018)