Çukurova Üniversitesi
Mühendislik Fakültesi Dergisi
Çukurova University
Journal of the Faculty of Engineering
ISSN: 2757-9255 CİLT/VOLUME: 39 SAYI/ISSUE: 3 EYLÜL/SEPTEMBER 2024

An Evidential Mask Transformer for Left Atrium Segmentation

Fatmatülzehra USLU [1,a]

[1]*Bursa Technical University, Department Electrical and Electronics Engineering, Bursa, Türkiye*

[a]***ORCID**: 0000-0001-7153-7583*

Article Info
Received : 05.01.2024
Accepted : 27.09.2024
DOI: 10.21605/cukurovaumfd.1560046

Corresponding Author
Fatmatülzehra USLU
fatmatülzehra.uslu@btu.edu.tr

Keywords
***M**ask transformer*
***I**mage segmentation*
***U**ncertainty*
***M**RI images*
***E**vidential learning*

***How to cite:** USLU, F., (2024). An Evidential Mask Transformer for Left Atrium Segmentation. Cukurova University, Journal of the Faculty of Engineering, 39(3), 639-646.*

ABSTRACT

The segmentation of the left atrium (LA) is required to calculate the clinical parameters of the LA, to identify diseases related to its remodeling. Generally, convolutional networks have been used for this task. However, their performance may be limited as a result of the use of local convolution operations for feature extraction. Also, such models usually need extra steps to provide uncertainty maps such as multiple forward passes for Monte Carlo dropouts or training multiple models for ensemble learning. To address these issues, we adapt mask transformers for LA segmentation which effectively use both local and global information, and train them with evidential learning to generate uncertainty maps from the learned Dirichlet distribution, with a single forward pass. We validated our approach on the STACOM 2013 dataset and found that our method can produce better segmentation performance than baseline models, and can identify locations our model's responses are not trustable.

Kanıtsal Maske Dönüştürücü Model ile Sol Kulakçık Bölütlemesi

Makale Bilgileri
Geliş : 05.01.2024
Kabul : 27.09.2024
DOI: 10.21605/cukurovaumfd.1560046

Sorumlu Yazar
Fatmatülzehra USLU
fatmatülzehra.uslu@btu.edu.tr

Anahtar Kelimeler
***M**aske dönüştürücü*
***G**örüntü bölütleme*
***B**elirsizlik*
***M**anyetik rezonans görüntüleri*
***K**anıtsal öğrenme*

***Atıf şekli:** USLU, F., (2024). An Evidential Mask Transformer for Left Atrium Segmentation. Cukurova University, Journal of the Faculty of Engineering, 39(3), 639-646.*

ÖZ

Sol kulakçığın yeniden şekillenmesine sebep olan hastalıklarının tanısının konulabilmesi için, sol kulakçığın bölütlenmesi gerekmektedir. Bu amaçla, genel olarak, konvolusyenel ağlar kullanılmaktadır. Fakat bu modellerin performansı, yerel hesaplama yapmaları nedeniyle düşük olabilir. Belirsizlik haritaları üretebilmeleri için, Monte Karlo dropout ya da çoklu model eğitimi (ensemble) gibi yaklaşımlara ihtiyaç duyulur. Bu problemleri gidermek için, yerel ve global bilgiyi bir arada kullanan, maske dönüştürücü modelleri, sol kulakçık bölütlenmesi için adapte ettik. Belirsizlik haritalarını elde etmek için de bu modeller, kanıtsal öğrenme ile eğitildi. Böylece, öğrenilen Dirichlet dağılımı kullanılarak, tek adımda belirsizlik haritaları elde edilebildi. Öne sürülen yaklaşım, STACOM 2013 veri setinde test edildi ve karşılaştırılan modellerden daha başarılı performans gösterdiği gözlemlendi. Üretilen belirsizlik haritalarının, modelin kararsız olduğu yerlerde yüksek belirsizlik gösterdiği gözlemlendi.

1. INTRODUCTION

The segmentation of the left atrium (LA) in Magnetic Resonance Imaging (MRI) images is necessary to extract clinical parameters, such as ejection fraction, volume and geometrical characteristics, to identify diseases related to the remodeling of the LA such as atrial fibrillation [1-3]. The low contrast of the LA in MRI images and its complicated shape makes it harder to segment with automatic tools. Recent work has reported promising results generated by deep networks on this task [1-3]. However, the task remains challenging when images to be segmented are collected from different vendors of MRI machines, leading to the problem known as the data-shift problem. Given the high confidence of deep networks even for their inaccurate results, wrongly segmented images reduce the trust of clinicians in such image analysis tools.

Uncertainty maps can give clues to clinicians to understand where a segmentation model has low confidence in its decision. Ensemble models or Monte Carlo dropout networks have been used for uncertainty map production [4]. The former approach trains multiple models to obtain a variety of segmentation masks, mimicking the behaviors of human experts with different medical expertise. On the other hand, the latter trains a single network with dropout layers, which are left active during test time. Monte Carlo simulation is performed with multiple runs of the same model, which produces multiple segmentation masks. Both approaches generate uncertainty maps by calculating variation across generated segmentation maps for the same input image. Despite their common use [4], these methods are costly, requiring training multiple models or running the same model multiple times to obtain different segmentation masks.

Recently, evidential learning (EL) was proposed for image classification [5] and later for segmentation problems [6-7]. This method estimates closed-form prediction distribution, based on the Dempster–Shafer Theory of Evidence [8]. The outputs of a network are described as categorical variables, and EL learns the parameters of prediction distributions over these variables, in contrast to yielding single point estimates posterior probabilities for each class output generated by the softmax function [8]. This property of EL makes it a good choice for model uncertainty calculation, which can be easily computed with a single forward pass from network outputs. Despite its high potential to improve the reliability of deep models, a few studies used EL for uncertainty calculation on medical image analysis, where the segmentation of brain tumors [6] and lymphomas [7] were examined. As far as we are aware, no previous study examined its use for uncertainty prediction of cardiac image segmentation.

Previous methods for LA segmentation generally used Convolutional Neural Networks (CNN) to learn local information in various abstraction levels [1-3]. Recent work introduced a new type of model enhancing CNN/transformer features with transformer decoders, called *mask transformers* [9], which was shown to outperform previous methods using CNN or only encoder-decoder transformers for semantic segmentation and object detection [9-10]. In these models, a pre-trained segmentation model provides local information to transformer decoders to learn global information in input images with self-attention and cross-attention mechanisms. Despite their high performance in image analysis [9-10], they struggle to reproduce small structures in input images. This may be due to yielding downsampled segmentation masks to reduce their high computational cost; for example, [9,11] produced segmentation masks four times smaller than the original size of input images. This property of mask transformers limits their applications to medical image analysis problems, where the segmentation of small structures is of high value [6-7].

In this study, we explore if the accuracy of LA segmentation and the reliability of its binary masks can be improved by training mask transformers with evidential loss. We modify the design of the mask transformers to increase their ability to segment small structures in input images. We present a training scheme specific to evidential learning to improve model performance, as well. Particularly, we investigate whether the two-steps training of mask transformers plays any role in its segmentation performance, which allows different model parts -- CNN and transformer decoders -- to be trained with different loss functions. We validated our method on a public LA segmentation dataset, the STACOM 2013 dataset [12]. We conduct ablation experiments to examine how our design choices for mask transformers change their segmentation performance. We also produce uncertainty maps to indicate locations where our model fails to yield confident results.

2. RELATED WORK

2.1. Evidential Learning

Evidential learning can make learning probability distributions over classes possible for a deterministic network. It uses the concepts of Dempster–Shafer Theory of Evidence [8], a generalization of Bayesian theory to subjective probability, where a belief distribution is obtained by assigning belief masses to the subsets of a discerning frame. The belief distribution for the M-classes classification problem can be represented with a Dirichlet distribution with M parameters, which are calculated from the outputs of the network, called *evidences*. Evidences are continuous values equal to or larger than zero so the activation function of the last layer of the network should be selected accordingly. For example, the softplus function can be used for this aim. The conversion between Dirichlet distribution parameters, α, and evidences, e, can be calculated with $\alpha_m = (e_m + 1)$ for the class m. Predicted classification probabilities for the class m are calculated with $\check{y}_m = \frac{\alpha_m}{S}$ for one-hot encoded class labels vector $\mathbf{y}$, where $S = \sum_{m=1}^{M} \alpha_m$ is the Dirichlet strength.

One can define an image segmentation task as an M-class classification problem for each pixel k in input images, and can minimize the Bayes risk of the cross-entropy loss to train the network with evidential learning; however, this loss calculation was found to be less stable than minimizing the Bayes risk of mean square loss [5], whose formula is given below:

$$E_k(\boldsymbol{\theta}) = \int \|\boldsymbol{y}_k - \check{\boldsymbol{y}}_k\|_2^2 \frac{1}{B(\boldsymbol{\alpha}_k)} \prod_{m=1}^{M} \check{y}_k^{\alpha_{km}-1} \mathrm{d}\check{\boldsymbol{y}}_k \quad [1]$$

where B is a multinomial beta function. $\boldsymbol{y}_k$ is the ground truth label for k^{th} pixel and $\check{\boldsymbol{y}}_k$ is its corresponding estimate by the network. Equation (1) can be simply rewritten in equation (2)

$$E_k(\boldsymbol{\theta}) = \sum_{m=1}^{M} (y_{km} - \check{y}_{km})^2 + \frac{\check{y}_{km}(1-\check{y}_{km})}{S_k+1} \quad [2]$$

2.2. LA Segmentation

Compared to other cardiac structures such as the left ventricle, there are few methods on the segmentation of the LA [1-3]. LA-Net [1] is a multi-task CNN model equipped with cross-attention and enhanced decoder modules to improve LA segmentation. TMS-Net [2] is a CNN ensemble model with an encoder and three decoders, which can segment the LA in MRI images along three orthogonal axes. The network also has a segmentation quality control module to eliminate poor segmentation masks. GSM-Net [3] is another CNN model, proposed to better use inter-slice similarities and the information at the temporal axis of CINE MRI images for LA segmentation, respectively, with a global slice sequence encoder and sequence-dependent channel attention module. These networks mostly used local information obtained with convolutional layers to segment the LA. However, they lack mechanisms that effectively learn global information in input images, which can limit their capacities when input images are noisy and have poor contrast.

3. METHOD

3.1. Evidential Learning for LA Segmentation

Firstly, we formulate the segmentation of the LA as a regression problem to minimize the Bayes risk of the sum of the square loss, given with equation (2). We use softplus function at the end of the proposed network to generate evidences, e, for each segmentation class, image background, and the LA. When calculating Dirichlet parameters, we use $\alpha = (e + 1)^2$ similar to [6], to easily reach high Dirichlet parameters, which increases the certainty of network outputs.

Similar to previous work [5-6], we assign pixel label predictions with misleading evidences to the uniform distribution, with Kullback-Leibler (KL) divergence loss, described with equation (3).

$$KL\,(D(\tilde{\boldsymbol{y}}_{\boldsymbol{k}}|\breve{\boldsymbol{\alpha}}_{\boldsymbol{k}})||D(\tilde{\boldsymbol{y}}_{\boldsymbol{k}}|\mathbf{1}) = log(\frac{\Gamma(\sum_{m=1}^{M}\breve{\alpha}_{km})}{\Gamma(\mathrm{M})\sum_{m=1}^{M}\Gamma(\breve{\alpha}_{km})}) + \sum_{m=1}^{M}(\breve{\alpha}_{km}-1)\,[(\psi(\breve{\alpha}_{km})\text{-}\psi(\sum_{m=1}^{M}\breve{\alpha}_{km})] \quad [3]$$

where $\Gamma(.)$ and $\psi(.)$ respectively denote the gamma and digamma function. $\breve{\boldsymbol{\alpha}}_k = \boldsymbol{y}_k + (\mathbf{1}-\boldsymbol{y}_k)\boldsymbol{\alpha}_k$ denotes the Dirichlet parameters for misleading evidences.

Therefore, our total loss function becomes as given below:

$$E(\boldsymbol{\theta}) + \lambda\, KL(D(\tilde{\boldsymbol{y}}|\breve{\boldsymbol{\alpha}})||D(\tilde{\boldsymbol{y}}|\mathbf{1}) \quad [4]$$

where we set $\lambda = 0.1\; min(1, t/5)$ for a current epoch of t.

3.2. Uncertainty Map Prediction

We use normalized entropy for uncertainty prediction, which can be calculated with $-\frac{1}{log(M)}\sum_{m=1}^{M}\breve{y}_m\, log\;\breve{y}_m$, for each pixel in an input image [6].

3.3. Our Segmentation Model

As shown in Figure 1, our segmentation model consists of three main parts: (i) a pixel encoder-decoder sub-network to generate pixel features, (ii) a transformer decoder stack to learn mask and class embeddings, and (iii) a segmentation mask prediction module. We will explain each part of our model, and the interaction between them, as follows.

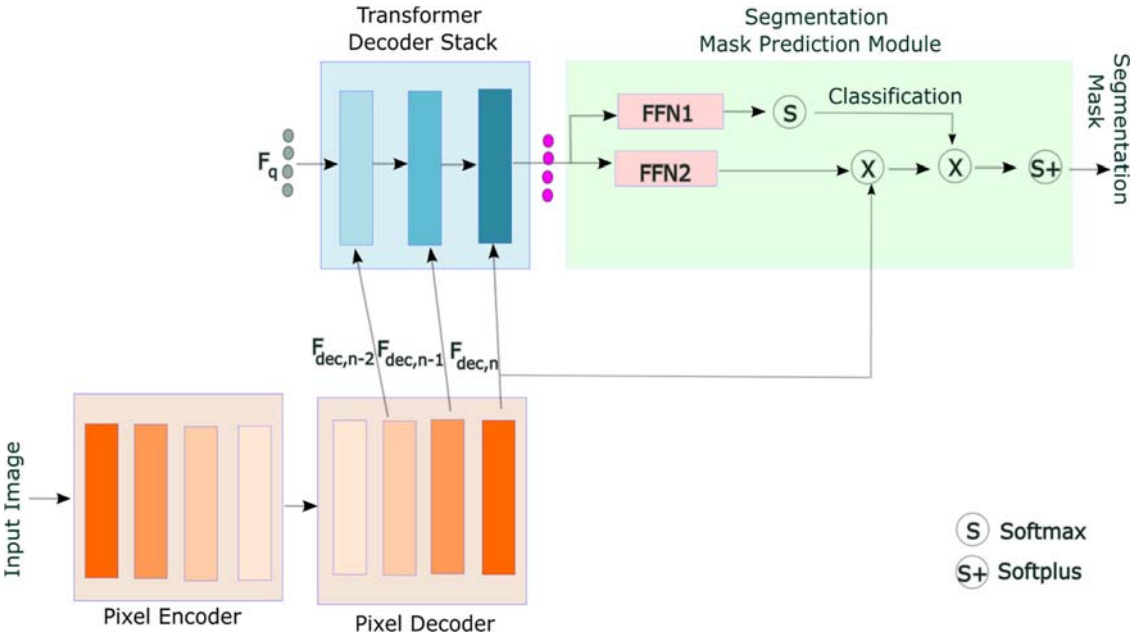

Figure 1. An overview of our network. FFN: feed forward network. Our model uses high resolution decoder features to enhance the segmentation of small structures in predicted masks

3.3.1. Pixel Encoder-Decoder Sub-network

We use Res-UNet [13] as a pixel encoder-decoder network, with five encoder layers and five decoder modules. Each module consists of two convolutional layers and a residual connection. Convolutional layers are followed by a group normalization layer and the parametric ReLU activation function, apart from the last decoder module, which has a convolutional layer with 2 channels followed by the softmax function to generate binary masks for image background and the LA classes.

After being fully trained for segmentation, its weights are frozen, and the outputs of its three decoder modules right before the final decoder module are used as pixel features, -- $F_{dec,n-2}$, $F_{dec,n-1}$ and $F_{dec,n}$ -- for the transformer decoder stack in our model (see Figure 1).

3.3.2. Transformer Decoder Stack

This module updates input query features, F_q, with pixel features through several decoder transformers, to generate mask and class embeddings in the segmentation mask prediction module [9] (see Figure 1). The initial query features, F_q, are evolved to be precursors of mask and class embeddings, $\tilde{F}_q$, with the cross-attention and self-attention mechanisms in each transformer decoder in the stack.

Figure 2 shows a detailed schematic of a transformer decoder [14]. Firstly, query features go through a linear projection to obtain $Q_n \in R^{N\times C}$, where N denotes the number of object masks and C represents the dimension of projected query feature vectors. Also, pixel features, generated by *pixel encoder-decoder module*, are linearly projected to yield keys and values matrices. For example, for $F_{dec,n}$, we produce $K_n \in R^{dim_n \times C}$ and $V_n \in R^{dim_n \times C}$, where dim_n corresponds to the dimension of projected decoder features. Then, the cross attention mechanism updates query features with $\tilde{F}_q = softmax(Q_n K_n^T)V_n + Q_n$.

In the transformer decoder stack, query features produced by a previous transformer decoder are sent to the next one in sequence to progressively update them. The stack in our model contains three transformer decoders, each being fed with a different scale of image features for cross attention mechanism, --$F_{dec,n-2} \in R^{256\times\frac{H}{4}\times\frac{W}{4}}$, $F_{dec,n-1} \in R^{128\times\frac{H}{2}\times\frac{W}{2}}$ and $F_{dec,n} \in R^{64\times H\times W}$--.

In contrast to previous work [9], we use higher resolution of pixel features to obtain segmentation maks, with the same size as input images. In order to reduce computation costs as a result of using high-resolution pixel features, we use a few object masks and reduce the dimensions of pixel features with linear projections to a small number such as 32.

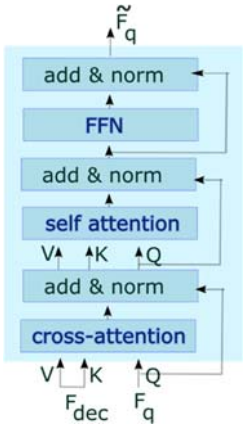

Figure 2. A transformer decoder module. $\boldsymbol{F_{dec}}$ and $\boldsymbol{F_q}$ respectively represent pixel and query features. K: keys, V: values and Q: queries

3.3.3. Segmentation Mask Prediction Module

This module consists of two feed-forward networks (FFN), and converts updated query features to class embeddings with $C_e = FFN_1(\tilde{F}_q)$ and mask embeddings with $M_e = FFN_2(\tilde{F}_q)$. Mask embeddings are multiplied with a highest resolution of pixel features, $F_{dec,n}$, to generate object masks with M=$M_e F_{dec,n}$, where $M \in R^{N\times H_n W_n}$ represents N object masks. Later, these object masks are transformed to segmentation masks by multiplying object masks with class embeddings $\tilde{M} = MC_e{}^T$, where $\tilde{M} \in R^{2\times HW}$ and $C_e \in R^{N\times 2}$. Note that, we learn object masks more than the number of classes in this setting.

In contrast to previous work [9,11], *our segmentation mask prediction module* uses image features with a resolution equal to that of the input image; this largely ensures the reproduction of small structures in segmentation masks.

4. RESULTS

4.1. Material

We assessed the performance of our model on the publicly available Stacom 2013 dataset [12]. The dataset contains MRI images obtained with balanced steady-state free precession (bSSFP) acquisition and has a resolution of 1.25 mm x 1.25 mm x 2.7 mm. The dataset consists of 10 MRI images for model training, and 20 MRI images for performance evaluation. The dataset provides binary masks for both the LA and proximal pulmonary veins. However, we combine them to generate a single mask for each image, similar to previous work [2].

4.2. Experiments

We used the ResUnet trained with the cross entropy (CE) and another ResUnet trained with evidential learning (EL) as our baseline models and compared their performance against our model.

4.2.1. Experiment 1

This experiment investigates the performance of our model for different loss functions. We examine three scenarios: (i) ResUnet, used in pixel encoder-decoder sub-network, and the rest of our model is trained with EL, (ii) ResUnet and the rest of our model are trained with the CE loss, (iii) the ResUnet is trained with CE and the rest of our model is trained with EL.

4.2.2. Experiment 2

This experiment examines the importance of using high-resolution features as pixel features, in the accuracy of produced segmentation masks. We prepare three different versions of pixel feature sets for our mask transformer. For a fair comparison with the original model, we use the same number of transformer decoders in all models, which is 3. We upsample the generated segmentation masks to ensure they have the same size as the ground truth segmentation masks when necessary.

Evaluated sets of pixel feature resolutions are (i) only rough resolution fetaures, -- $F_{dec,n-2} \in R^{256\times\frac{H}{4}\times\frac{W}{4}}$ --, for each transformer, (2) rough resolution features, -- $F_{dec,n-2} \in R^{256\times\frac{H}{4}\times\frac{W}{4}}$ --, for the first two transformers and moderate resolution fetaures, -- $F_{dec,n-1} \in R^{128\times\frac{H}{2}\times\frac{W}{2}}$ --, for the last one, (3) rough -- $F_{dec,n-2} \in R^{256\times\frac{H}{4}\times\frac{W}{4}}$ --, moderate -- $F_{dec,n-1} \in R^{128\times\frac{H}{2}\times\frac{W}{2}}$-- and fine resolution -- $F_{dec,n} \in R^{64\times H\times W}$ -- fetaures (as in the original model).

4.3. Experimental Setup

We first trained the Res-UNet for 100 epochs with a learning rate of 0.005, and the rest of the model was trained for 40 epochs with a learning rate of 0.005. We used a weight decay of 0.001 for the training of both models.

We used decoder features of the ResUnet with the spatial resolutions of 48x48 pixels, 96x96 pixels, and 192x192 pixels, as input to the transformer decoders respectively. The embeddings of keys, values, and queries have dimensions of 32. The number of queries was set to be 4. We added sinusoidal positional encodings to the pixel features, prior to feeding them to any transformer decoder.

4.4. Performance Metrics

We report Dice and Jaccard scores when measuring overlapped areas between the reference and predicted volumes. ASSD and HD distance are used when calculating distances between the boundaries of the two volumes.

4.5. Results and Discussion

Our model outperformed baseline models with large margins for three performance metrics, with a Dice score of 0.90, a Jaccard score of 0.82, and an ASSD of 1.85 (see Table 1). The baseline models are naively ResUnets trained with either CE or EL losses. On the other hand, our model consists of a transformer decoder stack and a segmentation mask prediction module, in addition to ResUnet used as pixel encoder and decoder. Despite its increased complexity compared to the baseline models, our model has a small overhead of less than 0,054 million parameters, in addition to the parameter of the Res-UNet, which is approximately 18 million parameters. This small overhead leads to a large performance improvement over baseline models. Another observation from Table 1 is that despite providing an easy uncertainty calculation, EL loss yields an underperforming ResUnet, compared to the CE loss.

A similar observation is made when training our model with CE and EL losses, as described in Experiment 1 in Section 4.2.1. Training our model purely with EL loss leads to poorer performance with a Dice score of 0.87 (see Table 2); however, its performance is still better than the baseline model of ResUnet trained with EL loss, which was 0.85. Training our model with purely CE loss outperforms the baseline model of ResUnet trained with CE loss, with a margin of 0.01. Finally, the best performance was obtained when the ResUnet was trained with the CE loss and the rest of our model with EL, as described in scenario 3, with a Dice score of 0.90 (see Table 2). These results show the effectiveness of our two-steps training scheme. It also shows that using a better-performing segmentation model --- ResUnet trained with CE instead of EL loss -- is important to reach a better performance by our model.

We also analyzed how much feature resolution is necessary for the transformer decoder stack, to obtain the most accurate segmentation masks. Table 3 summarises the results of Experiment 2 detailed in Section 4.2.2. We found that using higher-resolution pixel features in the transformer decoder stack leads to better segmentation performance, instead of repeating the same set of lower-resolution features such as $F_{dec,n-2}$. Incremental improvement is obtained by gradually increasing the resolution of features in the transformer decoder stack (see Table 3). The best segmentation performance is obtained when the finest resolution features $F_{dec,n}$ were included in the stack. This shows the importance of increasing the resolution of features in the transformer decoder stack to maintain small details in segmentation masks.

Figure 3 shows segmentation masks and uncertainty maps generated by our model. Our method produces very close responses to the ground truth masks and accurately reproduces small structures, as well as larger ones. Although deterministic networks are naively not capable of producing uncertainty maps, we generate the maps thanks to training our model with the EL loss. Similar to human-expert labeling, the uncertainty maps show higher uncertainty for the boundary of the LA. Boundary pixels are hard to label, and they are generally known to lead to high uncertainty even among medical experts [12].

5. CONCLUSIONS

As far as we are aware, it is the first time that mask transformers are used for the segmentation of cardiac MRI images. We modified the design of mask transformers to better segment small structures of the LA by incrementally increasing the resolution of decoder features in the transformer decoder stack. We trained our model with EL to generate uncertainty maps from the learned Dirichlet distribution over LA and image background classes. We also presented a training scheme to improve the segmentation performance of our model, where our two-steps training with CE and EL losses produced the best segmentation performance for the LA in MRI images.

We found that our model outperforms baseline models with large margins in the STACOM 2013 dataset. Another superiority of our method is that it can generate uncertainty maps with a single network and a single pass, therefore, it is less costly in terms of computation, and expected to work faster, compared to ensemble models and Monte Carlo dropout method. Future work will explore its use for the segmentation of other cardiac structures such as ventricles.

Table 1. Performance comparison of LA segmentation

Methods	Dice	Jaccard	HD	ASSD
ResUnet & **EL**	0.85 ±0.04	0.75 ±0.05	29.62 ±13.21	2.59 ±0.48
ResUnet & **CE**	0.88 ±0.03	0.79 ±0.05	**22.63** ±10.69	2.00 ±0.36
Our model	**0.90** ±0.03	**0.82** ±0.04	30.07±22.86	**1.85** ±0.42

Table 2. Performance comparison for three scenarios in Experiment 1

#Scenario	Dice	Jaccard	HD	ASSD
1	0.87 ±0.03	0.78 ±0.05	38.23 ±18.71	2.55 ±0.68
2	0.89 ±0.03	0.80 ±0.04	**29.24** ±23.03	1.90 ±0.36
3	**0.90** ±0.03	**0.82** ±0.04	30.07±22.86	**1.85** ±0.42

Table 3. Performance comparison for Experiment 2, where we use different combinations for pixel feature resolution. $\boldsymbol{F_{dec,n-2}} \in \boldsymbol{R}^{256\times\frac{H}{4}\times\frac{W}{4}}$, $\boldsymbol{F_{dec,n-1}} \in \boldsymbol{R}^{128\times\frac{H}{2}\times\frac{W}{2}}$, and $\boldsymbol{F_{dec,n}} \in \boldsymbol{R}^{64\times H\times W}$

Decoder features	Dice	Jaccard	HD	ASSD
$F_{dec,n-2}$ & $F_{dec,n-2}$ & $F_{dec,n-2}$	0.89±0.03	0.80±0.05	35.71±25.11	2.05±0.45
$F_{dec,n-2}$ & $F_{dec,n-2}$ & $F_{dec,n-1}$	0.89±0.03	0.81±0.04	32.68±24.28	1.94±0.41
$F_{dec,n-2}$ & $F_{dec,n-1}$ & $F_{dec,n}$	**0.90** ±0.03	**0.82**±0.04	**30.28** ±22.75	**1.86** ±0.42

6. REFERENCES

1. Uslu, F., Varela, M., Boniface, G., Mahenthran, T., Chubb, H., Bharath, A.A., 2021. LA-Net: a multi-task deep network for the segmentation of the left atrium. IEEE Transactions on Medical Imaging, 41(2), 456-464
2. Uslu, F., Bharath, A.A., 2023. TMS-Net: a segmentation network coupled with a run-time quality control method for robust cardiac image segmentation. Computers in Biology and Medicine, 152, 106422.
3. Uslu, F., 2023. GSM-Net: a global sequence modelling network for the segmentation of short axis CINE MRI images. Computerized Medical Imaging and Graphics, 102266.
4. Gawlikowski, J., Tassi, C.R.N., Ali, M., Lee, J., Humt, M., Feng, J., Zhu, X.X., 2023. A survey of uncertainty in deep neural networks. Artificial Intelligence Review, 56(Suppl 1), 1513-1589.
5. Sensoy, M., Kaplan, L., Kandemir, M., 2018. Evidential deep learning to quantify classification uncertainty. Advances in Neural Information Processing Systems, 31.
6. Li, H., Nan, Y., Del Ser, J., Yang, G., 2023. Region-based evidential deep learning to quantify uncertainty and improve robustness of brain tumor segmentation. Neural Computing and Applications, 35(30), 22071-22085.
7. Huang, L., Ruan, S., Decazes, P., Denœux, T., 2022. Lymphoma segmentation from 3D PET-CT images using a deep evidential network. International Journal of Approximate Reasoning, 149, 39-60.
8. Yager, R.R., Liu, L.(Eds.), 2008. Classic works of the dempster-shafer theory of belief functions. Springer, 219).
9. Cheng, B., Schwing, A., Kirillov, A., 2021. Per-pixel classification is not all you need for semantic segmentation. Advances in Neural Information Processing Systems, 34, 17864-17875.
10. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S., 2020. End-to-end object detection with transformers. In European Conference on Computer Vision (213-229). Cham: Springer International Publishing.
11. Yu, Q., Wang, H., Qiao, S., Collins, M., Zhu, Y., Adam, H., Chen, L.C., 2022. K-means mask transformer. In European Conference on Computer Vision (288-307). Cham: Springer Nature Switzerland.
12. Tobon-Gomez, C., Geers, A.J., Peters, J., Jürgen W., Karen, P., Rashed, K., et al., 2015. Left atrial segmentation challenge 2013: MRI testing. Figshare. Dataset.
13. Zhang, Z., Liu, Q., Wang, Y., 2018. Road extraction by deep residual u-net. IEEE Geoscience and Remote Sensing Letters, 15(5), 749-753.
14. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Polosukhin, I., 2017. Attention is all you need. Advances in Neural Information Processing Systems, 30.