Anatomy-Aware Adaptive Feature Perturbation Framework for Semi-Supervised MRI Segmentation

Supplementary Material

6. The Proposed Semi-supervised MRI Image Segmentation Framework

As shown in Fig. 5, the main framework contains the preprocessing module, four U-shaped networks in parallel and two different supervision means for labelled images and unlabelled images. The preprocessing module firstly randomly rotates, flips and resizes the labelled images to 256 $\times$ 256 as data augmentation, the weak and strong perturbation process will be described in the next section. The four networks have the same structure and parameters as U-Net except for the middle one with an additional AAAFPM. The labelled images and weak perturbated images are collectively fed into U-Net to generate the pseudo labels of weak perturbated images for the feature-level perturbation network and two strong perturbated images in each iteration according to the shared weight derived by the supervision of ground truth for labelled images.

6.1. Preliminaries

The total objective function L consists of two parts: supervised loss L_s and unsupervised loss L_{us}. Supervised loss has the same weight as unsupervised loss in the objective function, as shown:

$$L = \frac{1}{B}\sum_{n=1}^{B}(\frac{1}{2}(L_s + L_{us})), \tag{4}$$

B is the batch size. L_s calculates the cross-entropy loss and dice loss between the predictions P_i and ground truth y_i, as shown as:

$$L_s = \frac{1}{2}(L_{cross\_entropy}(p_i, y_i) + L_{dice}(p_i, y_i)), \tag{5}$$

where $p_i = f_\theta(x_i)$,f_θ, x_i denotes the model under the parameter set θ and model's i_{th} input , the θ is learnable.

L_{us} comprises the three loss function L_{SP_1}, L_{SP_2} and L_{FP}, as shown:

$$L_{us} = \alpha L_{FP} + \beta L_{SP_1} + \gamma L_{SP_2}, \tag{6}$$

where $\alpha + \beta + \gamma = 1$, Inspired by contrastive learning [23], UniMatch uses two strong image-level perturbations to fully exploit the original perturbation, L_{SP_1}, L_{SP_2} represents the dice loss between two strong perturbated images' predictions p^U_{s1}, p^U_{s2} with weak perturbated images' prediction p^U_w, as shown:

$$\begin{aligned} L_{SP_1} &= \sum 1(max(p^U_w) \geq \tau) L_{dice}(p^U_{s1}, C_1(p^U_w)), \\ L_{SP_2} &= \sum 1(max(p^U_w) \geq \tau) L_{dice}(p^U_{s2}, C_2(p^U_w)), \end{aligned} \tag{7}$$

In the equation, $1()$ denotes the indicator function and τ is the pre-defined threshold to get the high-confidence pseudo label, $p^U_{s1} = f_{\theta_U}(x^U_{s1})$, $p^U_{s2} = f_{\theta_U}(x^U_{s2})$, f_{θ_U} is the model under the parameter set θ_U, $x^U_{s1} = P_{s1}(x^U_w)$, $x^U_{s2} = P_{s2}(x^U_w)$, $P_{s1}()$ and $P_{s2}()$ additionally uses Colour Jitter, Gaussian blur and CutMix as strong perturbations compared with weak perturbations. Notably, P_{s1} and P_{s2} are independent and unequal. $x^U_w = P_w(x^U)$, $p^U_w = f_{\theta_L}(x^U_w)$, x^U and $P_w()$ are the unlabelled images and the weak perturbations such as rotation. For the supervision of strong perturbated images' predictions, the CutMix operation $C_1()$ and $C_2()$ on the weak perturbated images' predictions are the same as that used in the strong perturbation.

Besides, UniMatch also applies feature-level perturbations , L_{FP} represents the dice loss between feature perturbated images' predictions p^U_{fp} with weak perturbated images p^U_w, as shown:

$$L_{FP} = \sum 1(max(p^U_w) \geq \tau) L_{dice}(p^U_{fp}, p^U_w), \tag{8}$$

$p^U_{fp} = f_{\theta_{LFP}}(x^U_w)$, $f_{\theta_{LFP}}()$ denotes the model within dropout under the parameter set θ_{LFP}.

6.2. The Basic Units

The classic U-Net is used as the backbone to reduce the performance gained from network design, and it has some modifications compared with the original one. For simplicity, the ConvBlock in the encoders and final convolution in the decoders are ignored As illustrated in Fig. 5.

The encoder of U-Net includes one ConvBlock and four DownBlocks. The ConvBlock consists of two sets of convolution, batch normalization, and Leaky ReLU with a dropout placed between them. The kernel size and strides are 3$\times$ 3 and 1 for convolution operation. Followed by ConvBlock, there are four consecutive DownBlocks, each DownBlock comprises max pooling and CovBlock. The introduction of dropout reduces the risk of overfitting caused by the small amount of available labelled data in semi-supervised semantic segmentation, and the setting of dropout for the encoder is [0.05,0.1,0.2,0.3,0.5].

The decoder of U-Net consists of four UpBlocks and a convolution. In UpBlock, the nearest upsampling increases the resolution of the features at the current stage to that of the previous stage, the upsampled features are concatenated with those at the current stage, and are fed into ConvBlock mentioned above which has the same parameters without dropout. After four consecutive UpBlocks, the final con-

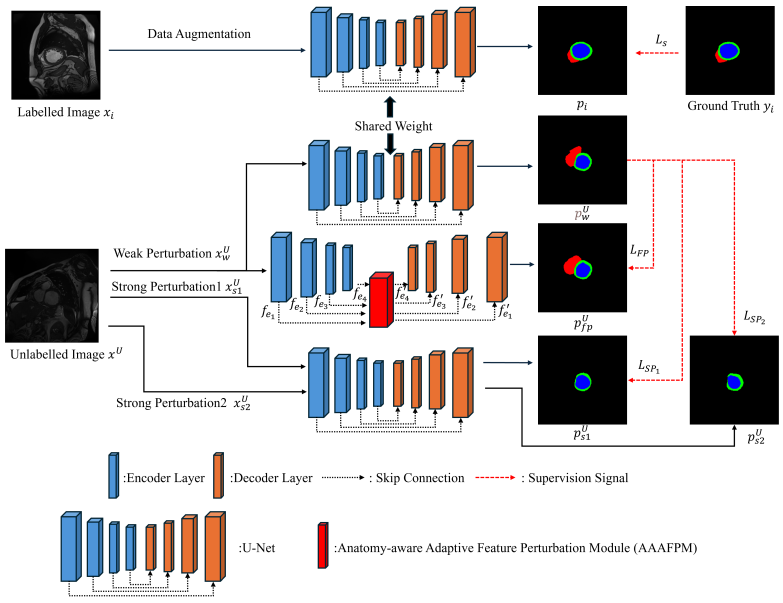


Figure 5. The proposed semi-supervised MRI image segmentation framework. The cardiac MRI images are shown as examples of medical images. The four networks have the same structure and parameters as U-Net except for the middle one with an additional AAAFPM. For simplicity, the ConvBlock in the encoders and final convolution in the decoders are ignored. The labelled images are augmented and fed into the network and its predictions are supervised by the ground truth. Then the weight is shared to generate the predictions for weak perturbated images, the weak perturbated image's predictions are the pseudo labels for the feature-level perturbation networks and two kinds of strong perturbated images. The framework not only contains one weak and two strong image-level perturbations but also includes one feature-level perturbation, which is more task-specific.

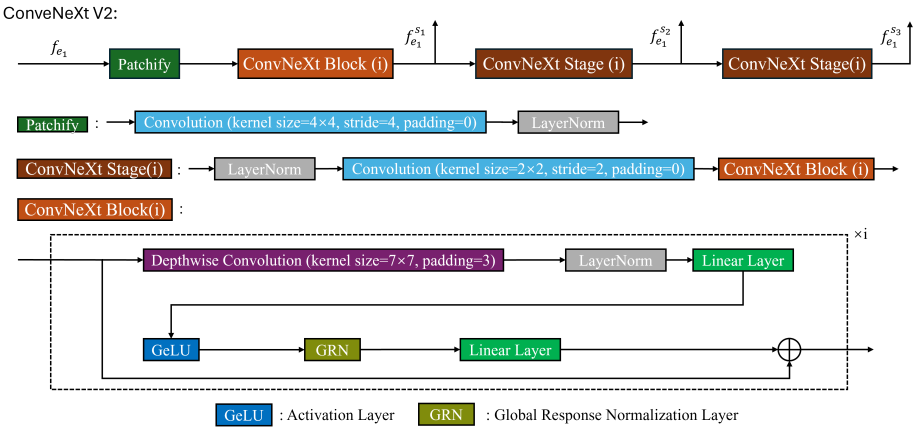


Figure 6. The modified ConvNeXt V2. Compared with the original one, the output of each stage is exploited to the following information process. Besides, the original dropout module in each block is removed to facilitate the analysis of the proposed module's improvement.

volution with 3× 3 kernel size and 1 stride processes the decoded features to make the predictions.

7. Ablation Studies

Tab. 6 demonstrates the performance of AAAFPF at the second stage under the different number of labelled data, the number in the bracket in the first column denotes the maximum value of the standard deviation of Gaussian noise. For example, 'noise =0.05' denotes the maximum value of the standard deviation of Gaussian noise is 0.05. Tab. 6 shows that the value of the maximum standard deviation of the Gaussian noise depends on the amount of labelled data. The proportion of labelled data determines the necessity of consistency regularisation. It becomes essential when the labelled data amounts to 1% or 5%, because the model needs stronger Gaussian noise intensity to effectively

Table 6. The performance of AAAFPF with different maximum intensities of Gaussian Noise under different numbers of labelled data

Method	Labelled Data	DSC (%) ↑				JS (%)↑				95HD (mm)↓				ASD (mm)↓			
		RV	MC	LV	Avg	RV	MC	LV	Avg	RV	MC	LV	Avg	RV	MC	LV	Avg
AAAFPF(noise=0.05)	1%	84.48	83.45	91.90	86.61	74.15	72.05	85.40	77.20	2.01	1.50	3.30	2.27	0.63	0.60	0.71	0.65
AAAFPF(noise=0.1)		83.31	85.55	91.57	86.81	72.94	74.92	84.95	77.60	2.37	1.99	2.19	2.18	0.80	0.58	0.69	0.69
AAAFPF(noise=0.2)		**86.06**	**86.62**	92.14	**88.27**	**76.23**	**76.57**	86.01	**79.60**	**1.69**	1.23	2.16	1.69	**0.52**	0.47	0.52	0.50
AAAFPF(noise=0.5)		83.08	86.52	**92.27**	87.29	72.64	76.45	**86.24**	78.44	2.39	**1.16**	**1.26**	**1.60**	0.75	**0.36**	**0.34**	**0.48**
AAAFPF(noise=0.05)	5%	89.08	86.91	92.69	89.56	80.92	77.08	86.99	81.66	1.45	1.26	1.92	1.54	0.40	0.46	0.46	0.44
AAAFPF(noise=0.1)		88.94	86.59	92.36	89.30	80.75	76.59	86.30	81.21	1.53	2.89	1.92	2.11	0.41	0.62	0.43	0.49
AAAFPF(noise=0.2)		89.14	86.65	**93.32**	89.70	80.94	76.67	**87.79**	81.80	**1.37**	1.97	1.93	1.76	**0.33**	0.44	0.53	0.43
AAAFPF(noise=0.5)		**89.33**	**87.86**	93.24	**90.14**	**81.27**	**78.53**	87.74	**82.51**	1.54	**1.13**	**1.14**	**1.27**	0.43	**0.30**	**0.26**	**0.33**
AAAFPF(noise=0.05)	10%	90.46	**88.87**	**94.25**	**91.19**	83.05	**80.16**	**89.39**	**84.20**	**1.23**	**1.07**	**1.81**	**1.37**	**0.32**	**0.30**	**0.43**	**0.35**
AAAFPF(noise=0.1)		**90.69**	88.37	93.36	90.81	**83.50**	79.39	88.00	83.63	1.57	2.84	2.99	2.47	0.39	0.62	0.92	0.64
AAAFPF(noise=0.2)		89.60	88.35	93.87	90.61	81.78	79.38	88.76	83.31	1.35	1.88	2.33	1.85	0.39	0.51	0.61	0.50
AAAFPF(noise=0.5)		90.15	88.54	93.48	90.72	82.67	79.66	88.11	83.48	1.47	1.83	3.46	2.25	0.38	0.51	0.97	0.62
AAAFPF(noise=0.05)	20%	**91.50**	89.04	94.43	91.66	84.72	80.48	89.72	84.97	**1.15**	2.99	3.77	2.64	0.39	0.59	0.80	0.59
AAAFPF(noise=0.1)		90.66	88.85	93.90	91.14	83.53	80.17	88.81	84.17	1.38	1.37	2.22	1.66	0.36	0.42	0.57	0.45
AAAFPF(noise=0.2)		91.11	88.94	94.35	91.47	84.16	80.32	89.53	84.67	1.24	1.72	1.91	1.62	0.34	0.36	0.51	0.40
AAAFPF(noise=0.5)		91.48	**89.09**	**94.82**	**91.80**	**84.73**	**80.55**	**90.38**	**85.22**	1.24	**1.04**	**1.06**	**1.11**	**0.33**	**0.29**	**0.29**	**0.30**

Table 7. Impact of dropout at different stages on performance under different number of labelled data

Method	Labelled Data	DSC (%) ↑				JS (%)↑				95HD (mm)↓				ASD (mm)↓			
		RV	MC	LV	Avg	RV	MC	LV	Avg	RV	MC	LV	Avg	RV	MC	LV	Avg
BWoD(L2)	5%	**89.50**	**87.41**	92.59	**89.83**	**81.76**	**77.84**	86.56	**82.05**	1.48	**1.96**	2.91	2.12	0.40	0.53	0.73	0.55
BWoD(L3)		88.65	87.12	92.90	89.56	80.22	77.39	**87.20**	81.60	**1.36**	2.04	**1.37**	**1.59**	**0.38**	0.52	**0.43**	**0.44**
BWoD(L4)		88.24	86.72	92.84	89.27	79.63	76.77	87.03	81.14	1.61	2.12	3.09	2.27	0.43	0.65	1.09	0.72
BWoD(L5)		87.44	87.11	**92.79**	89.11	78.36	77.37	87.08	80.94	1.60	1.97	2.16	1.91	0.49	**0.51**	0.54	0.51
BWoD(L2)	10%	90.19	88.46	93.77	**90.81**	82.64	79.55	88.50	**83.56**	1.39	1.88	3.20	2.16	0.38	0.49	0.76	0.54
BWoD(L3)		87.85	**88.81**	93.92	90.19	79.07	**80.07**	88.83	82.66	2.85	1.88	2.69	2.47	0.89	0.44	0.70	0.68
BWoD(L4)		**90.24**	88.10	93.34	90.56	**82.80**	79.03	87.88	83.24	**1.36**	2.02	2.22	1.87	**0.35**	0.55	0.62	0.51
BWoD(L5)		89.63	88.53	**94.05**	90.74	81.69	79.63	**89.04**	83.45	2.34	**1.12**	**1.58**	**1.68**	0.50	**0.39**	**0.34**	**0.41**

use unlabelled data. The model depends on labelled data to learn features when it receives strong supervision from 10% labelled data, so it needs low noise intensity to maintain stable supervision from labelled data. The model achieves stable performance when the proportion of labelled data exceeds 20%. Besides, the model becomes more robust to unlabelled data and less prone to overfitting, so a larger noise intensity is reintroduced at this stage.

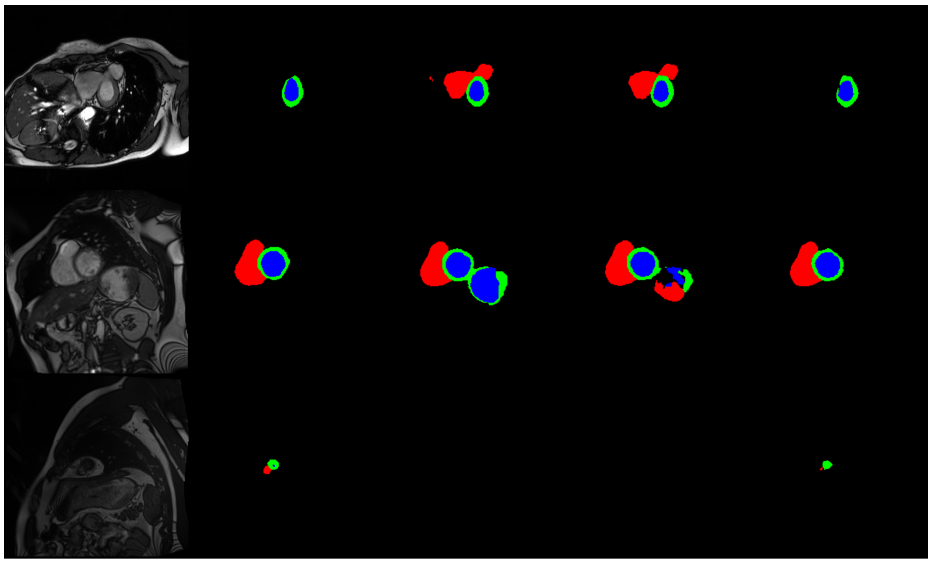
Furthermore, BWoD(L2), BWoD(L3), BWoD(L4), BWoD(L5) in Tab. 7 represents removing channel dropout at the 2nd, 3rd, 4th and 5th stage, it turns out that removing the dropout operation at the second stage is the best choice especially using 5% labelled data, because the feature map at the second skip connection contains the anatomical information, and the random dropout will destroy the information and degrade performance.

Besides, in Tab. 8, AAASFPM-S3 and AAASFPM-S4 denote that the proposed AAASFPM is placed at the third and fourth skip connections, respectively, and dropout operations are implemented in the other skip connections. The proposed AAASFPM placed at the second skip connections is denoted as AAAFPF, which outperforms that of other skip connections using 1%, 5% and 10% labelled data. It shows that our method aims at analysing low-level and detailed information at the second skip connection rather than the high-level and abstract information at the third and fourth skip connections.

Fig. 7 and Fig. 8 show some visual results under 5% and 10% labelled images respectively. The introduction of Gaussian noise reduces the baseline segmentation errors as shown in the first row of both Fig. 7 and Fig. 8. However, the fixed Gaussian noise intensity brings extra noise during the pseudo-label generation and disturbs the discriminative feature learning, resulting in the background misclassifica-

Table 8. The performance of AAASFPM at the different locations under different numbers of labelled data

Method	Labelled Data	DSC (%) ↑				JS (%)↑				95HD (mm)↓				ASD (mm)↓			
		RV	MC	LV	Avg	RV	MC	LV	Avg	RV	MC	LV	Avg	RV	MC	LV	Avg
AAAFPF	1%	**86.06**	**86.62**	92.14	**88.27**	**76.23**	**76.57**	86.01	**79.60**	**1.69**	**1.23**	2.16	**1.69**	**0.52**	**0.47**	**0.52**	**0.50**
AAASFPM-S3		83.29	85.38	91.38	86.69	72.86	74.71	84.71	77.43	2.48	2.06	2.71	2.42	0.83	0.58	0.70	0.70
AAASFPM-S4		82.63	86.31	**92.46**	87.13	72.07	76.09	**86.33**	78.16	5.12	2.96	**2.13**	3.40	1.44	0.80	0.55	0.93
AAAFPF	5%	89.33	**87.86**	93.24	**90.14**	81.27	**78.53**	**87.74**	**82.51**	1.54	**1.13**	**1.14**	**1.27**	0.43	**0.30**	**0.26**	**0.33**
AAASFPM-S3		89.49	87.13	92.58	89.73	**81.54**	77.44	86.66	81.88	**1.32**	2.01	2.31	1.88	**0.35**	0.54	0.64	0.51
AAASFPM-S4		**89.50**	87.07	**93.29**	89.95	81.50	77.32	87.72	82.18	1.41	1.95	2.52	1.96	0.44	0.59	0.55	0.53
AAAFPF	10%	90.46	**88.87**	**94.25**	**91.19**	83.05	**80.16**	**89.39**	**84.20**	**1.23**	**1.07**	**1.81**	**1.37**	**0.32**	**0.30**	**0.43**	**0.35**
AAASFPM-S3		90.32	88.25	92.94	90.51	82.86	79.18	87.35	83.13	1.33	2.93	3.57	2.61	0.37	0.65	1.09	0.70
AAASFPM-S4		**90.58**	88.53	93.91	91.01	**83.25**	79.63	88.78	83.89	2.75	1.92	2.50	2.39	0.63	0.53	0.59	0.59
AAAFPF	20%	91.48	89.09	**94.82**	91.80	84.73	80.55	**90.38**	85.22	1.24	**1.04**	1.06	1.11	0.33	**0.29**	0.29	**0.30**
AAASFPM-S3		**91.60**	**89.29**	94.81	**91.90**	84.88	**80.89**	90.31	**85.36**	**1.17**	1.07	**1.05**	**1.10**	0.30	0.32	**0.28**	**0.30**
AAASFPM-S4		91.53	88.62	93.01	91.05	**84.91**	79.84	87.48	84.07	1.23	2.76	3.32	2.44	**0.28**	0.57	1.10	0.65



(a) MRI Image (b) Ground truth (c) Baseline (d) BGN (e) AAAFPF

Figure 7. Results of ablation studies (5% labelled data). The performance of baseline, BGN and AAAFPF is compared in the MRI image. Red, green and blue colours represent right ventricle, myocardium and left ventricle correspondingly. Compared with other methods, AAAFPF makes the best prediction that matches well with the ground truth.

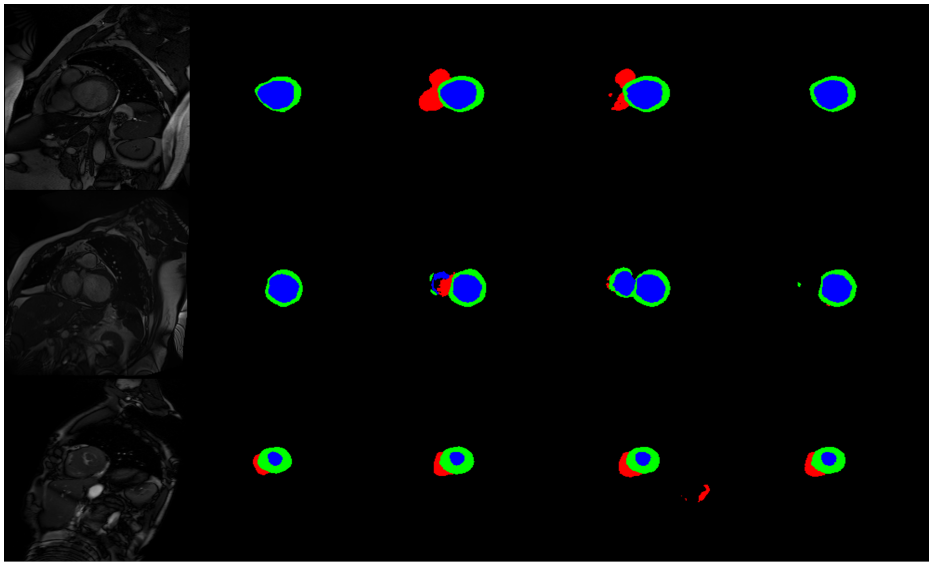
tion like MC and LV as shown in the second rows of Fig. 8. The proposed AAAFPF learn the descriptive and discriminative features and maintain the robustness by dynamically adjusting the intensity of Gaussian noise further reducing the BGN's errors. Consequently, AAAFPF makes the best prediction that matches well with the ground truth.

8. Experiments

As illustrated in Fig. 9, the proposed AAAFPF demonstrates superior performance in segmenting the right ventricle, which outperforms other methods that struggle with this task. Moreover, the proposed approach accurately delineates both the left ventricle and the myocardium. However, occasionally the background misclassification as the right ventricle is observed.

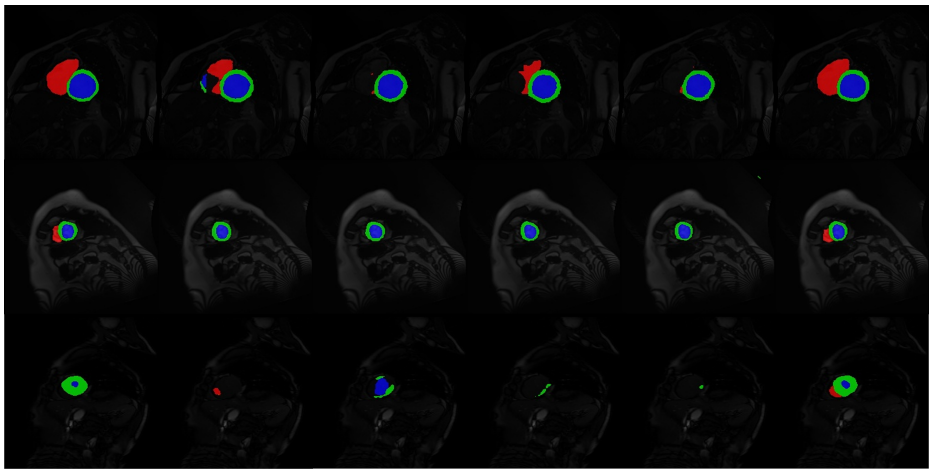
9. Discussion

The proposed method is based on a weak-to-strong consistency framework (baseline), which uses pseudo-labels as supervised signals and leverages consistency regularization to improve performance. This framework exceeds the consistency regularisation method such as CCT with 6.47% in-



(a) MRI Image (b) Ground truth (c) Baseline (d) BGN (e) AAAFPF

Figure 8. Results of ablation studies (10% labelled data). The display order and the legend of each image are the same as that of Fig. 7. AAAFPF reduces the errors caused by the fixed intensity of Gaussian noise and makes more correct predictions.



(a) Ground truth (b) CCT (c) CPS (d) UAMT (e) URPC (f) AAAFPF

Figure 9. A few example of segmentation by CCT, CPS, UAMT and URPC and the proposed method in ACDC dataset.The legend of each image are the same as that of Fig. 7. The proposed method makes better predictions in terms of the right ventricle, left ventricle and myocardium compared with other methods.

crease and 4.16mm drop in DSC and 95HD respectively, and the pseudo-label method such as CPS with 5.09mm and 1.59mm decreases in 95HD and ASD as illustrated in Tab. 4 and Tab. 1.

Additionally, the AAASFPM plays a critical role in the framework, which aggregates contextual features from multiple stages of ConvNeXt V2 to construct hypercolumn features for richer semantic representation. Pixel-level similarity is calculated by enriched semantic information that allows the model to dynamically emphasise significant areas rather than processing all features equally for enhanced feature representation. Notably, the introduction of AAASFPM improves performance across all metrics under relatively small labelled data, such as 5% and 10% labelled data, as illustrated in the first four rows of Tab. 5.

The AAAFPF further improves the model by discriminative feature learning. In detail, the AAASFPM dynamically adjusts Gaussian noise intensity to enhance pseudo-label reliability based on local variation and global structure constraints. Besides, the auxiliary CFPM expands the per-

turbation space to enrich feature diversity. Consequently, it avoided the common mistakes made by channel-level perturbations like dropout and the straightforward combination of channel-level and spatial-level perturbations like Gaussian noise. For example, AAAFPF has segmentation results closely aligned with ground truth, while other methods misclassify background regions as cardiac structures (e.g. right ventricle, myocardium, and left ventricle), as demonstrated in Fig. 8.

The AAAFPF effectively combines perturbations from both image and feature levels, which outperforms approaches that rely on single perturbation. For instance, AAAFPF achieves significantly lower delineation errors compared to single perturbations and more precise segmentation even for small objects, as shown in the second row of Fig. 9 and the last row of Fig. 2.