Stylization-Based Image Segmentation: Comparative Analysis of RAIN and AdaConv for Scar Tissue Assessment in LGE Magnetic Resonance Images

Mohammad Shanur Rahman[1], Mingxuan Gu[1]

[1]Fakultät für Pattern Recognition, FAU Erlangen-Nürnberg
mohammad.shanof.rahman@fau.de

Abstract. Accurate segmentation of Magnetic Resonance Images (MRIs) is crucial for assessing scar tissue extent in myocardial infarction diagnosis. Utilizing multiple modalities enhances diagnostic accuracy and enables comprehensive assessment by providing complementary anatomical and tissue structure information. However, domain shift occurs between modalities, rendering segmentation models trained on one modality unfit for others. Unsupervised Domain Adaption (UDA) serves as a bridge between modalities by learning styles from unlabeled target domain data and adapting labeled source domain data accordingly. This work compares two domain adaptation methods, Random Adaptive Instance Normalization (RAIN) and Adaptive Convolutions (AdaConv), for Structure-Aware Style Transfer. These methods are popular for UDA due to their ability to align the style of images across different domains without requiring paired training data or explicit annotations, thereby enabling the transfer of knowledge and visual characteristics from a source domain to a target domain in an unsupervised manner. Our findings show that RAIN exhibits better generalization across inter-modality styles, primarily due to the adversarial styled training of the segmentation model. Additionally, we demonstrate that AdaConv alters the structure of the endocardial and epicardial contour in source images, resulting in significant boundary misalignment despite higher pixel-wise overlap with the ground truth. Nevertheless, performance improves when AdaConv's pre-training is conducted using images from the target modality.

1 Introduction

Myocardial infarction (MI) is a significant contributor to global mortality rates [1]. Among diagnostic tools, cardiovascular magnetic resonance (CMR) stands out for its ability to provide a comprehensive, multifaceted view of the heart. Different CMR sequences, such as Later Gadolinium Enhancement CMR (LGE), T2-weighted CMR (T2), and balanced-Steady State Free Precession (bSSFP), offer complementary information that can be integrated to generate a final anatomical and functional depiction of the heart [2]. LGE provides anatomical and functional information of the heart, and is very helpful in visualizing MI. T2 captures the acute injury and ischemic regions, while bSSFP captures cardiac motions and presents clear boundaries [3]. Fig. 1 provides an example of the bSSFP and LGE sequences. LGE enhances the infarcted myocardium, to appear distinctively bright, compared with the healthy tissues, and therefore is effective in determining the presence, location, and extent of MI. In addition, images captured from multiple modalities of the same subject have the potential to facilitate the segmentation of structures whose boundaries may not be fully distinguishable in any of the images[2].

As quantified infarct size represents a crucial endpoint in clinical trials [1], it is imperative to delineate the ventricles and myocardium from LGE CMR images. Nonetheless, manual segmentation of these structures is a time-consuming and tedious task, prone to inter- and intra-observer variability. Thus, automated segmentation methods are highly desirable in clinical practice [2].

Medical imaging often faces the challenge of limited labeled data availability. To address this issue, deep models can be trained on a large scale with training set images from a dissimilar modality. But since, as shown in Fig. 1(b), there can be considerable amount of shift in data distribution across two different modalities, the data used for inference can be significantly different from the one used for training [4]. Nevertheless, training the segmentation model with images from different modalities can be challenging, particularly when domain labels are not readily available. In this context, unsupervised domain adaptation can help reduce the gap between a source and target domain by stylizing the labeled source domain data to the unlabeled target domain data.

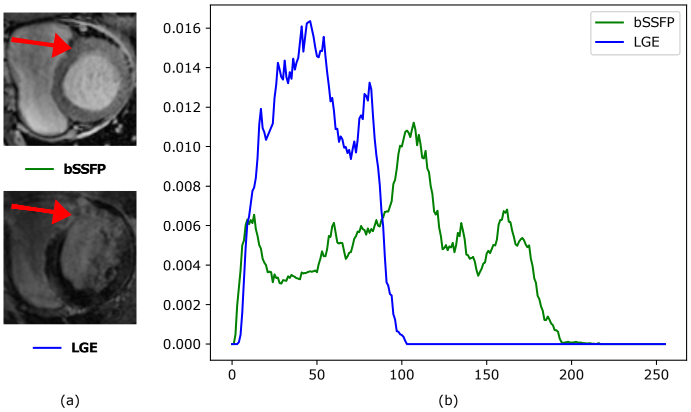


Fig. 1. The differences of image appearance (a) and intensity distributions (b)in the cardiac region (the union of LV, MYO, RV) between LGE images and bSSFP images.

In this work, we focus on training image segmentation models using LGE images, which provides better information about scar tissue but lacks labeled data. To address this limitation, we employ a technique of stylizing the labeled bSSFP images to resemble LGE images. Subsequently, we train the segmentation model on the newly generated synthetic LGE-stylized labeled images. For the purpose of stylization, we present comparative analysis of two different approaches: Random Adaptive Instance Normalization (RAIN), which mainly captures the global statistics and Adaptive Convolutions (AdaConv), which is able to transfer local structures and statistics.

Our findings reveal that AdaConv changes the structure in stylized cardiac images, which leads to mis-alignment of images with their ground truth labels, and hence the performance of segmentation model based on AdaConv performs poorly as compared to RAIN, which maintains the global structure after stylization. Furthermore, we demonstrate the comparable performance in terms of pixel-wise overlap of AdaConv based segmentation model when pre-training was conducted using images from target modality. Which shows that AdaConv is comparatively better in transferring styles within modalities as compared to transferring styles across modalities. However, due to the structural modifications introduced by AdaConv, their is still a significant boundary misalignment in the predictions made.

2 Related Work

The segmentation network for LGE images is trained even in the absence of labeled LGE data by leveraging anatomical knowledge and features learned from easily annotatable and acquirable bSSFP images and unlabelled LGE images [5]. In this context, unsupervised domain adaptation techniques have been extensively explored and can be broadly classified into two categories: 1) feature-level adaptation with latent space alignment; 2) pixel-level adaptation with image-to-image translation.

2.1 Feature level adaption

This approach focuses on aligning the distributions between domains in the latent space by minimizing measures of distance between features extracted from the source and target domains. The source and target inputs are mapped into a shared latent feature space, enabling a classifier learned from this common space to work for both domains. For instance, Kamnitsas et al. [6] proposed an early attempt to align feature distributions in cross-protocol MRI images using adversarial loss. Zhang et al. [7] proposed multi-view adversarial training for dataset-invariant left and right-ventricular coverage estimation in cardiac MRI.

2.2 Image level adaption

In this approach, images from the target domain are transformed to resemble the appearance of the source domain, enabling segmentation models trained on the source domain to be used for target images, and vice versa. For example, a multi-modal image translation network was utilized to generate synthetic LGE images from a single annotated bSSFP image [8]. In this work, image-to-image translation from one domain to another was achieved by feeding the decoder of one domain with the content of its own domain but the style of another domain. Additionally, a two-stage approach was proposed in another work [9], where CT images were first translated to appear like MRI using CycleGAN, and then both the generated MRI and a few real MRI were used for semi-supervised tumor segmentation.

3 Contribution

We do a comparative analysis of two image level adaption techniques, namely RAIN and AdaConv for segmentation in CMR images. Our main contributions include -

- Implementations of RAIN and AdaConv based style transfer for unsupervised domain adaption across modalities of CMR images.
- Comparative analysis of DRU-Net based segmentation model trained on stylized images gernerated using RAIN and AdaConv.
- Comparison of performance of segmentation models in three different settings as summarised in Table 1.

Tab. 1. The table summarizes the experiment setups, where Setting 0 serves as the baseline without any stylization. The subsequent settings differ in the choice of modality used for training the stylization module during pretraining and segmentation.

	Pretraining		**Segmentation**	
	Source Domain	**Target Domain**	**Source Domain**	**Target Domain**
Setting 0	None	None	bSSFP	None
Setting 1	bSSFP	T2	bSSFP	Random LGE Image
Setting 2	bSSFP	LGE	bSSFP	Random LGE Image
Setting 3	bSSFP	LGE	bSSFP	Corresponding LGE Image

4 Methods

4.1 Dataset

The STACOM MS-CMRSeg 2019 challenge dataset, consisting of short-axis cardiac MR images from 45 patients diagnosed with cardiomyopathy, was employed in this study. The dataset was obtained from Shanghai Renji Hospital with institutional ethics approval [10]. Each patient underwent CMR imaging using three sequences, namely late gadolinium enhancement (LGE), T2-weighted (T2), and balanced steady-state free precession (bSSFP). Ground truth masks delineating cardiac structures, including the left ventricle cavity (LV), right ventricle cavity (RV), and myocardium of the left ventricle (Myo), were provided for 40 training samples (T2 and bSSFP), 40 test samples (LGE) and 5 validation samples. Affine transformations, such as scale, rotation, translation, and shear, were applied to augment the dataset. Subsequently, the sequences were center-cropped to 224 x 224 pixels for RAIN to retain the region-of-interest (ROI), and for AdaConv the sequences were center-cropped to 256 x 256 pixels, because it uses a fully connected layer that restricts input matrices to be of fixed dimension (3, 256, 256).

4.2 Domain Adaption

The objective is to train a 4-class pixel-wise segmentation model on a set of unlabeled images $D_{LGE}(x_t)$. To achieve this, another labeled modality called $D_{bSSFP}(x_s, y_s)$ is used. Where x_s are images and y_s are labels. However, since training a model on one domain and using it for inference on another domain yields very poor results [11], the x_s from D_{bSSFP} are stylized to have content of bSSFP modality but style of LGE. By synthesizing these new images, which have the same content as the labeled images, a segmentation model can be trained on them. And since, the images have styles closer to target modality, the problem of domain shift is solved.

For stylization, we have utilized two modules - RAIN[12] and AdaConv[13]. These modules were trained using x_s from D_{bSSFP} as the source image and x_s from D_{T2} as the target style image, which is different from using LGE images as the target style during actual segmentation. This approach is more realistic, as images from target domain might not always be available. And we are also able to use available T2 images in our segmentation pipeline.

4.2.1 RAIN Module. It is based on Adaptive Instance Normalization (AdaIN)[14], which primarily aims to align the mean and variance of the content feature with those of the style features. It generates stylized images which have the appearance of the style image while preserving the structure of the content images. This is accomplished by the use of an encoder-decoder architecture. Initially, a VGG encoder is utilized to encode both the content and style images. In the feature space, the style transfer is performed by applying the AdaIN layer, which is governed by (1). Following this, a decoder is learned to invert the AdaIN output back to the image space.

$$\text{AdaIN}(f_c, f_s) = \sigma(f_s)\frac{f_c - \mu(f_c)}{\sigma(f_c)} + \mu(f_s) \tag{1}$$

where f_c,f_s are the latent features of the content and style image, $\mu()$ and $\sigma()$ are used to perform the channel-wise mean and standard deviation. The VGG encoder used to extract the content and style features is used again to compute the content loss L_c and style loss L_s. The content loss L_c is the Euclidean distance between the target features and the features of the output image. And style loss L_s matches the mean and variance of style features extracted by VGG encoder with mean and variance of target style features.

Fig. 3 shows the iterative search of RAIN for new stylized images that go beyond the target samples, thereby improving the adaptation performance of the task-specific module. It makes the style searching process a differentiable operation, enabling an end-to-end style search through gradient back-propagation. This is achieved by introducing a style variational auto-encoder (VAE) between the encoder and the decoder modules of AdaIN.

The style-VAE as presented in Fig. 2 includes E_{vae} and a decoder D_{vae}. E_{vae} encodes the style $\mu(f_s) \bigoplus \sigma(f_s)$ (where $\bigoplus$ denotes concatenation) to a Gaussian distribution $N(\psi, \xi)$. And D_{vae} decodes a sampling ϵ from the distribution to reconstruct the original style. So, apart from L_c and L_s of AdaIN, RAIN has two more losses i.e. L_{KL} and L_{Rec}. L_{KL} is the KL divergence of $N(\psi, \xi)$ with respect to $N(0, 1)$ and L_{Rec} is the reconstruction loss between $\mu(f_s) \bigoplus \sigma(f_s)$ and $\overline{\mu(f_s) \bigoplus \sigma(f_s)}$, where $\overline{\mu(f_s) \bigoplus \sigma(f_s)}$ is the reconstructed style vector from a sampling $\epsilon \sim N(\psi, \xi)$. Hence, the overall training objective for RAIN is to minimize the following loss:

$$L_{RAIN} = L_c + \lambda_1 L_s + \lambda_2 L_{KL} + \lambda_3 L_{REC} \tag{2}$$

where, λ_1 is the style loss weight, λ_2 is KL Divergence loss weight and λ_3 is weight for reconstruction loss. Further definition of losses in equation (2) are:

$$L_c = \|f(g(t)) - t\|_2 \tag{3}$$

$$L_s = \sum_{i=1}^{L} \|\mu(\phi_i(g(t))) - \mu(\phi_i(s))\|_2 + \sum_{i=1}^{L} \|\sigma(\phi_i(g(t))) - \sigma(\phi_i(s))\|_2 \tag{4}$$

$$L_{KL} = \mathrm{KL}[N(\psi, \xi)||N(0, I)] \tag{5}$$

$$L_{REC} = \|\mu(f_s) \bigoplus \sigma(f_s), \overline{\mu(f_s) \bigoplus \sigma(f_s)}\|_2 \tag{6}$$

f, g and t in equation(3) are the encoder, decoder and output of AdaIN respectively. Each ϕ_i in equation(4) denotes a layer in VGG-19 used to compute the style loss.

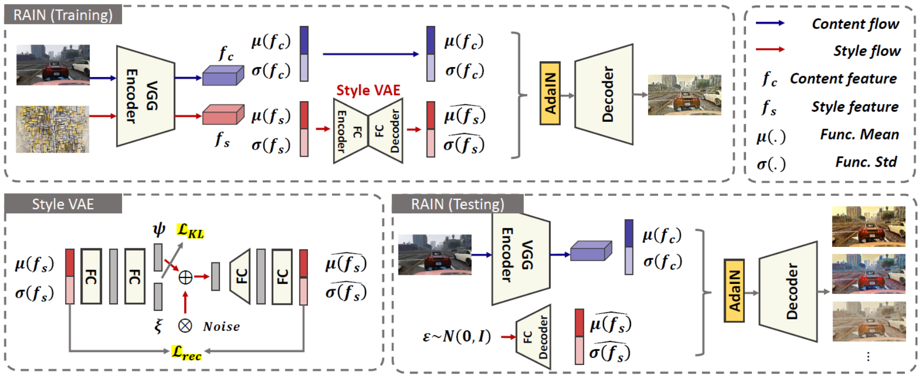


Fig. 2. Architecture of RAIN with it's extra VAE in the latent space to encode the style $\mu(f_s)$ and $\sigma(f_s)$ into a standard distribution. While testing, RAIN enables to generate newer styles using the ϵ drawn from the distribution learnt using training [12].

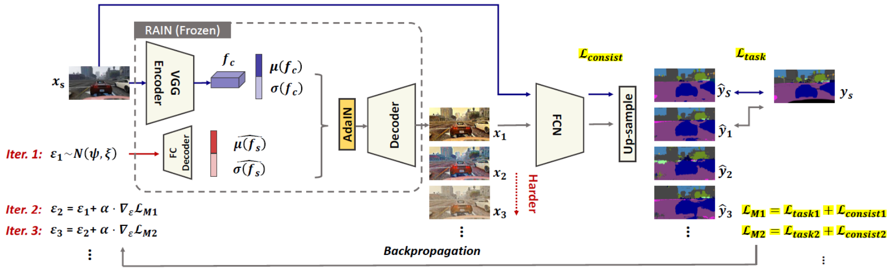

RAIN (Frozen)
x_s
VGG Encoder
f_c
$\mu(f_c)$
$\sigma(f_c)$
Iter. 1: $\varepsilon_1 \sim N(\psi, \xi)$
FC Decoder
$\widehat{\mu(f_s)}$
$\widehat{\sigma(f_s)}$
AdaIN
Decoder
x_1
x_2
x_3
Harder
FCN
Up-sample
$\mathcal{L}_{consist}$
$\mathcal{L}_{task}$
$\hat{y}_s$
y_s
$\hat{y}_1$
$\hat{y}_2$
$\hat{y}_3$
$\mathcal{L}_{M1} = \mathcal{L}_{task1} + \mathcal{L}_{consist1}$
$\mathcal{L}_{M2} = \mathcal{L}_{task2} + \mathcal{L}_{consist2}$
Iter. 2: $\varepsilon_2 = \varepsilon_1 + \alpha \cdot \nabla_\varepsilon \mathcal{L}_{M1}$
Iter. 3: $\varepsilon_3 = \varepsilon_2 + \alpha \cdot \nabla_\varepsilon \mathcal{L}_{M2}$
Backpropagation

Fig. 3. ϵ-based style exploration using RAIN. As the ϵ is increased in the positive direction of gradient, the segmentor gets even challenging images [12].

4.2.2 AdaConv Module. It is also an extension of AdaIN which proposes that the style of an image can be defined not only by its global statistical features, but also by its local structures and statistics [13]. Instead of transferring the global statistics (mean and standard deviation) from each style feature, AdaConv predicts full convolution kernels and bias values as transformations from the style image. These predicted kernels are then convolved with the features of the content image. As the kernels capture the localized spatial structure in the style, AdaConv is expected to accurately transfer the structural elements of the style image to the content image.

To extend AdaIN, AdaConv replaces the scale term (mean in equation 2) with a conditioning 2D style filter f, and adds a separable, pointwise convolution tensor p to the input style parameter.

For an input feature channel with values $x \in \mathbb{R}$, equation 1 can be re-written using target style represented with $\{a, b\}$, where a is mean and b is the standard deviation of style image features,

$$\mathrm{AdaIN}(x; a, b) = a\left(\frac{x-\mu_x}{\sigma_x}\right) + b \tag{7}$$

A conditioning 2D style filter $\mathbf{f} \in \mathbb{R}^{k_h X k_w}$ replaces a in equation (7). This filter helps in varying the feature channel in a spatial way, as per the local structure in a neighborhood $N(x)$ around pixel x,

$$\begin{aligned}\mathrm{AdaConv}_{dw}(x; \mathbf{f}, b) &= \sum_{x_i \in N(x)} f_i \left(\frac{x_i-\mu_x}{\sigma_x}\right) + b,\\ &= \sum_{x_i \in N(x)} \mathrm{AdaIN}(x; f_i, b)\end{aligned} \tag{8}$$

where $\mathrm{AdaConv}_{dw}$ is the depthwise AdaConv variant. And then, this depthwise variant in equation(8) is further extended by expanding the input style parameters to also include a separable, pointwise convolution tensor $\mathbf{p} \in \mathbf{R}^C$ for input with C feature channels. This enables AdaConv to capture style based on correlations across features x_c in different input channel c,

$$\mathrm{AdaConv}(x; \mathbf{p}, \mathbf{f}, b) = \sum_c \mathrm{AdaConv}_{dw}(x_c; \mathbf{f_c}, b_c). \tag{9}$$

In essence, style features encoded by AdaConv $\{\mathbf{p},\ \mathbf{f},\ \mathrm{b}\}$ includes a depthwise-separable 3D kernel, with depthwise and pointwise convolution components, and biases for each channel.

For style transfer using AdaConv, as presented in Fig. 4, input style and content image is encoded using pre-trained VGG-19 encoder to latent features S and C respectively. To predict the kernel, the style features S are further encoded by a style encoder E_s to obtain a global style descriptor W. And then, multiple kernels K_i are predicted for different resolutions of decoded image. Convolutions are done using the corresponding predicted kernels and the layers of decoder D, which later outputs the stylized image. The encoder E_s, the kernel predictors K_i and the decoder D are trained together to minimize the sum of content and style losses with λ as style loss weight using:

$$L_{AdaConv} = L_c + \lambda L_s \tag{10}$$

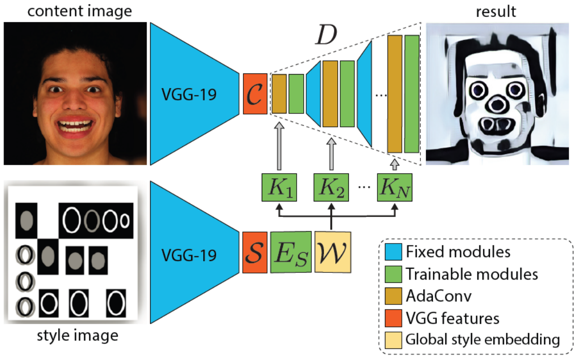


Fig. 4. Architecture of AdaConv with it's kernel predictors, for structure aware style transfer[13].

4.3 Experiment Setup

We use the setup of training segmentation model using images from D_{bSSFP} and using it for inference on D_{LGE} images as the baseline model. We further devised three additional scenarios. Wherein, we first pretrain the stylization module using images from either D_{T2} or D_{LGE}. And then, the stylised images which share the same label as D_{bSSFP} are used for training the DRU-Net based segmentation model. Table 1 provides a summary of the four experimental settings that were employed. The following section briefs about the motivation for each of the scenarios.

4.3.1 Scarce target domain images (*Setting 1*). This setting involves working with a setup where images from the desired target domain, such as LGE, are not available at all. To address this, a related modality, such as T2, may be used to pretrain the stylization module. This scenario closely resembles real-world conditions and is therefore one of the most difficult to handle. Since, the stylization module, while its trainig, never sees the actual target LGE style, there must be a way of exploring around T2 styles and learning more complex styles.

4.3.2 Abundant target domain images (*Setting 2*). In this situation, we make the assumption that abundant images from the target domain is available. To use this abundance, we first pretrain the stylization module using bSSFP as content and LGE as target style images. Once this pretraining is complete, we assume that the stylization module has learned the generalized target style transformations for LGE modality images. Therefore, any arbitrary LGE image can be fed as the target during segmentation. This approach is particularly well-suited for stylization modules that may not possess the capacity to explore complex styles, but are proficient at learning styles that are common to images within a given modality.

4.3.3 Modality based image registration (*Setting 3*). We build upon Setting 2 by incorporating the patient's corresponding LGE image as the style during segmentation. This approach is motivated by the fact that LGE and bSSFP images can provide complementary information, which when combined can enhance the accuracy of the segmentation process [2]. In this setting, the segmentation module is trained on images that share content with the bSSFP modality image, but style from the corresponding LGE modality image. This approach bears similarity to the process of image registration, although it cannot be guaranteed that both images were taken at the same cardiac position as required in registration. But since, the fact that both images correspond to the same patient motivates greater coherence and meaning to the resulting stylized image.

4.4 Segmentation Module: DRU-Net

The segmentation in stylized images are done using DRU-Net[15]. We chose DRU-Net as our segmentation architecture because it is specifically designed for image segmentation tasks in medical imaging applications. Its dual pathway architecture enables it to capture both local and global features in an input image, while its residual connections improve the model's training efficiency and performance. With state-of-the-art performance [15], DRU-Net has been a powerful and efficient network for image segmentation tasks in medical imaging applications.

4.5 Evaluation Metrics

4.5.1 Dice Coefficient. The Dice coefficient, denoted by $DC(A, B)$, is a widely used similarity metric to evaluate the performance of image segmentation algorithms. It measures the overlap between two sets of pixels, A and B, where A represents the set of pixels in the ground truth label for segmentation, and B represents the set of pixels predicted by the model for that class. The Dice coefficient is computed as the ratio of the size of the intersection of A and B to the size of their union. The formula for the Dice coefficient is given by:

$$Dice(A, B) = \frac{2|A \cap B|}{|A| + |B|}, \tag{11}$$

where $|.|$ represents cardinality operator, $|A|$ and $|B|$ represent the number of pixels in sets A and B, respectively, and $|A \cap B|$ represents the number of pixels that are in both sets. The Dice coefficient ranges from 0 to 1, with a value of 1 indicating a perfect match between the ground truth labels and the model's predictions, and a value of 0 indicating no overlap between them.

4.5.2 Hausdorff Distance. The Hausdorff distance, denoted by $HD(A, B)$, is a measure of the dissimilarity between two sets of points in a metric space. In the context of image segmentation, it is often used to measure the similarity between the object boundary of ground truth labels and the object boundary of predicted labels. The directed Hausdorff distance from the ground truth boundary A to the predicted boundary B is defined as the maximum distance of a point in A to its nearest neighbor in B, and is given by:

$$H(A, B) = \max_{a \in A} \min_{b \in B} |a - b|, \tag{12}$$

where $|\cdot|$ represents the Euclidean distance between points. DC and HD are often used in conjunction because of the complementary information they provide. DC measures the overlap between the segmented object and ground truth object, while HD captures the spatial discrepancy between their object boundaries. The DC provides an overall measure of segmentation performance, while HD can identify specific regions of disagreement.

5 Results and Discussion

5.1 Quantitative Comparison

Tab. 2. Overview of the segmentation performance of the DRU-Net model when trained on bSSFP images stylized to target LGE images under various settings. The evaluation was conducted using images from the LGE modality. The results demonstrate - i) Using AdaConv led to changes in tissue boundary structure, resulting in significant boundary misalignment between predicted segmentation and ground truth (High HD), despite achieving comparable pixel-wise overlap (High DC). ii) The model trained on AdaConv images showed improvement when presented with LGE images, indicating its limited ability to learn styles across different modalities. On the other hand, models trained on RAIN images demonstrated consistent performance, even in Setting 1 where T2 images were used for pretraining. iii) The performance of RAIN remains almost consistent across the settings, signaling a saturation in complexity of images generated during adversarial epsilon iterations (5 for this observation).

		DC ↑				**HD ↓**			
Setting	**Stylizer**	**MYO**	**LV**	**RV**	**AVG ± STD**	**MYO**	**LV**	**RV**	**AVG ± STD**
0	None	0.159	0.295	0.132	0.195±0.071	82.519	74.884	74.243	77.215±3.759
1	AdaConv	0.469	0.601	0.576	**0.549±0.057**	138.668	122.387	89.616	***116.890±20.399***
1	RAIN	0.573	0.846	0.714	**0.711±0.111**	6.499	9.927	10.469	8.965±1.758
2	AdaConv	0.536	0.777	0.733	**0.682±0.105**	105.094	105.788	110.827	***107.236±2.555***
2	RAIN	0.564	0.838	0.699	**0.700±0.112**	11.255	40.275	9.668	20.399±14.069
3	AdaConv	0.586	0.789	0.74	**0.705±0.086**	126.616	129.443	45.656	***100.572±38.848***
3	RAIN	0.594	0.825	0.758	**0.726±0.097**	6.747	12.612	12.636	10.665±2.77

As summarised in Table 2, it was observed that AdaConv changes the boundary structure in generated stylised image, because of which inspite of comparable Dice Coefficient, models trained on AdaConv-stylized images performs poorly in terms of Hausdroff Distance across all the three settings. The fact that AdaConv-based model performs poorer in Setting 1, but better in Settings 2 and 3 where, it was showed images from target modality shows it's inability to generalise across inter-modality styles. Hence, it is evident that AdaConv falls short in transferring styles to the extent that it can learn LGE-like styles from T2 images alone for accurate segmentation of LGE images.

On the other hand, RAIN incorporates epsilon sampling in the latent space. By tuning this epsilon in the positive direction of the gradient, the segmentation model can explore even more challenging and previously unseen images. It can be reasonably assumed

Tab. 3. Performance improvement in RAIN on increasing epsilon iterations.

	DC ↑				**HD ↓**			
Epsilon iterations	**MYO**	**LV**	**RV**	**AVG**	**MYO**	**LV**	**RV**	**AVG**
1	0.376	0.701	0.568	**0.548**	30.465	61.962	40.112	44.18
3	0.489	0.771	0.609	**0.623**	27.394	49.28	28.171	34.948
5	0.573	0.846	0.714	**0.711**	6.499	9.927	10.469	8.965

that this sampling-based style exploration contributes to the superior performance of the RAIN-based segmentation model on images with a newer style, specifically those from the LGE modality. And since, this style exploration is completely based on global lightning, there is no change in the structure, and models based on LGE have significantly lower HD compared to the models based on AdaConv. Figure 5 and Table 3 demonstrates that as the number of epsilon iterations increases (up to 5 iterations), the images become progressively more challenging, resulting in a stronger segmentation model.

Table 2 also illustrates an overall performance enhancement in Setting 3 when compared to Settings 2 and 1. As evident, the performance of RAIN exhibit limited responsiveness to changes in settings, with even a decrease in performance observed in Setting 2. This decrease can be attributed to RAIN reaching a saturation point in it's style exploration because of 5 epsilon iterations.

On the other hand, AdaConv demonstrates significant improvement in performance when LGE images are used during its pre-training process in Setting 2 and 3. This signals, if AdaConv is also equipped with the mechanism of exploring newer styles, by incorporating RAIN like epsilon in it's latent space, it can also provide even challenging images to the segmentation module for it's performance improvement. Although there is a need for a loss function which caters to changing structure of generated images.

5.2 Qualitative Comparison

Fig. 6 provides clear evidence that AdaConv exhibits poor performance in Setting 1 (Difference in relative peak position and count). However, when exposed to LGE images

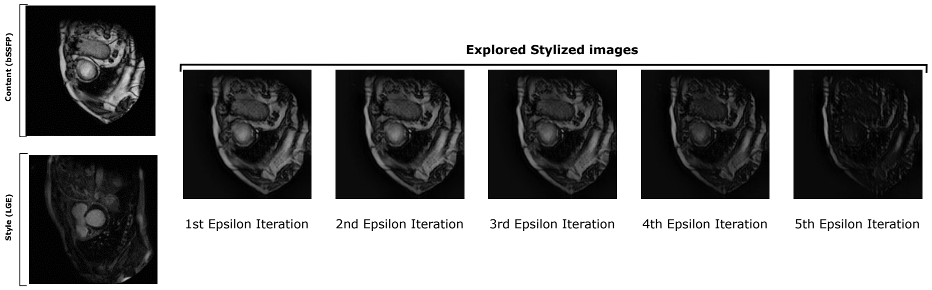


Fig. 5. Exploration of styles by RAIN module. As the epsilon iterations are increased, the segmentation module is fed by even challenging stylized images, thereby making it stronger.

during its pre-training in Setting 2, it demonstrates improved outcomes. This indicates that AdaConv is weak in transferring styles across modality.

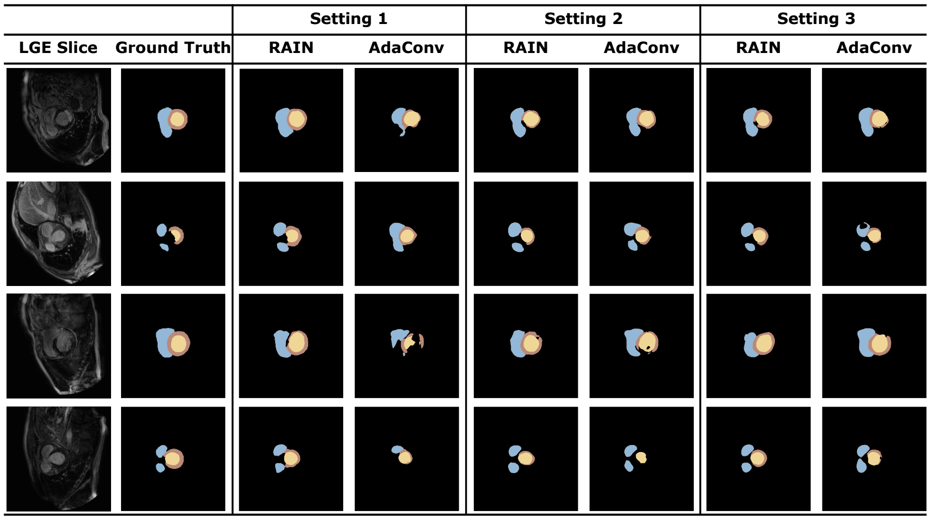
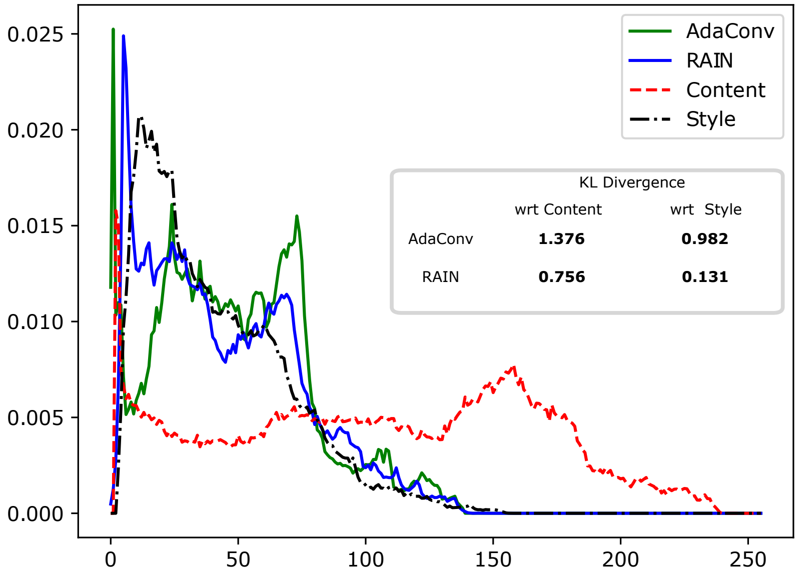


Fig. 6. Comparative performance of segmentation modules pre-trained with AdaConv and RAIN images in various settings, compared to Ground Truth. The evaluation was done on validation set of images from LGE modality. AdaConv performs poorly in Setting 1, and shows generally good segmentation results in Settings 2 and 3, which suggests of it's inability to learn inter-modality styles like RAIN.

Another valuable metric for comparison is the analysis of intensity distributions in stylized images produced by both modules. Fig. 8 illustrates this comparison, revealing that the histograms of myocaridum section in AdaConv's stylized images starts resembling those of RAIN from Setting 2 onwards. This shows that, AdaConv is unable to learn the inter-modality style. Fig. 7 shows that, images generated by RAIN are closer to both source content and target style in terms of intensity distribution as compared to those generated by AdaConv. Moreover, visible extra peaks are observed in AdaConv's intensity distribution, which heavily signals the change in local structures. This change heavily hampers the segmentation performance of AdaConv-based model.

6 Conclusion

In this work, we explored two different style transfer-based domain adaptation techniques for myocardial infarction segmentation in LGE images. The necessity for domain adaptation arises from the difficulty of manually labeling LGE images due to their low intensity delineation. Consequently, there is a scarcity or absence of labeled LGE images available to train a supervised segmentation model. To address this challenge, we trained the segmentation module on bSSFP images. However, since it is impractical to train on



KL Divergence		
	wrt Content	wrt Style
AdaConv	**1.376**	**0.982**
RAIN	**0.756**	**0.131**

Fig. 7. Comparison of Intensity distribution of RAIN and AdaConv stylized images with source content image and with target style image. The KL (Kullback-Leibler) divergences were calculated with respect to content as well as style images. It can be observed that intensity distribution of RAIN-stylized image is closer to both source content image and target style image.

one modality and perform inference on another, we employed stylization techniques to make the bSSFP images resemble the style of LGE images. Subsequently, a supervised model was trained on these stylized synthetic images.

By comparing the performance of the segmentation module using images stylized with RAIN and AdaConv, we observed that AdaConv based segementation model was performing poorly. To evaluate its further effectiveness, we devised two additional scenarios for pre-training and segmentation. In one scenario, we used LGE images as the target style during both pre-training and segmentation, while in the other scenario, we employed the patient's corresponding LGE image as the style during segmentation. The results showed a significant improvement in the segmentation model based on AdaConv-stylized images. These findings led us to conclude that AdaConv's style exploration capabilities are not as extensive as those of RAIN with five epsilon iterations. Moreover, the enhanced performance suggests that equipping AdaConv with a gradient-based style exploration mechanism would likely yield a notable improvement in the segmentation model's performance.

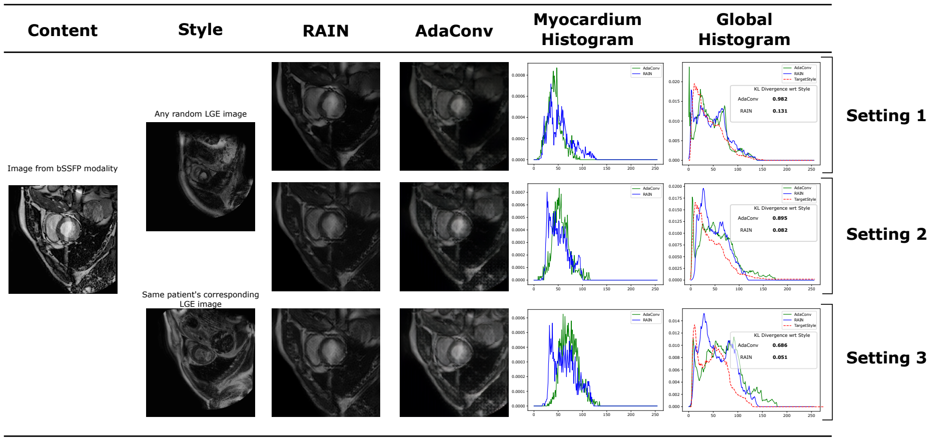


Fig. 8. Stylized Images generated by RAIN and AdaConv in three Settings. Intensity distribution comparison of AdaConv and RAIN for myocardium and union of RV, LV, and Myo sections. The KL divergence with respect to style decreases as we move towards Setting 3. It can be seen that, both AdaConv and RAIN intensity distributions follow the target style closely. But, AdaConv has some local peaks, which signals appearnce of newer structures with different intensities from content image.

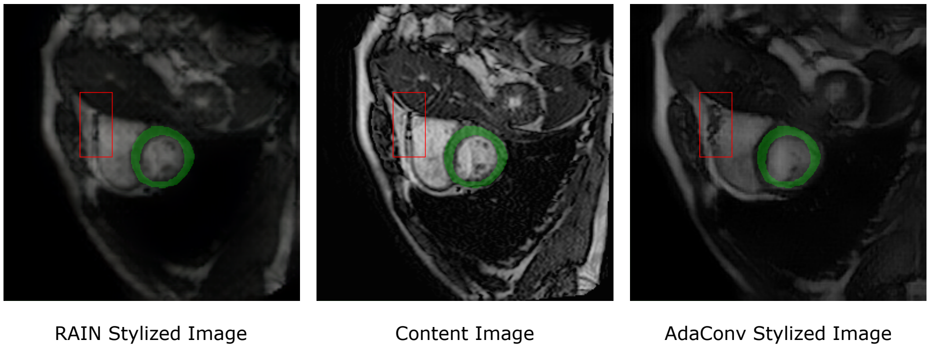

RAIN Stylized Image
Content Image
AdaConv Stylized Image

Fig. 9. Qualitative comparison of AdaConv and RAIN stylized images. The green highlighted circular sections represent the ground truth marking of the endocardium. It can be seen that RAIN (left) maintains the original structure of the endocardium as depicted by the content image (center), while in the AdaConv stylized image (right), the endocardium appears dilated and deviated from its intended structure. Additionally, the section highlighted with a red rectangle appears more exaggerated in the AdaConv stylized image compared to RAIN. These observations strongly indicate that AdaConv alters the local structure heavily during the stylization process, which if not aligned with the pixel-wise labels, results into boundary misalignment (high HD).

7 Future Work

We observed that, AdaConv alters the structure while stylization, because of which models trained on AdaConv stylized images have misaligned boundaries with ground truth as summarised in Table 2. And, we also observed that AdaConv's performance is poor when T2 images were used for pre-training as evident from Table 2.

For the first problem, in order to lower the effect of local structural transfers, visible in Fig. 9, we should increase the predicted kernel-size for convolutions. Thereby, making the style transfer still a convolution based transformation but more global.

Once, it is evident that AdaConv does not makes significant structural change, we can proceed to enable AdaConv to generalise for LGE images, just by showing it T2 images. This has to be accomplished by enabling it to explore newer styles. This exploring capability is already there in RAIN, where the ϵ sampled from the distribution in latent space of it's style-VAE is increased in the opposite direction of gradient, and hence more difficult image is available to the segmentor. To incorporate this style exploration in AdaConv, we should have a Variational Autoencoder (VAE) in it's architecture, then the ϵ sampled from latent distribution of VAE should be increased in the same direction of gradient. This way, during multiple such ϵ-based iterations, the segmentation model will get even challenging images to be trained on.

By addressing these issues and incorporating the proposed modifications, we can enhance AdaConv's performance and improve its ability to maintain structure yet transfer local styles and generalize across different image types.

References

1. Yusuf S, Hawken S, Ounpuu S, Dans T, Avezum A, Lanas F et al. Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): Case-control study. Lancet 364 (2004), pp. 937–52.
2. Zhuang X. Multivariate mixture model for myocardium segmentation combining multi-source images. CoRR abs/1612.08820 (2016).
3. Kim HW, Farzaneh-Far A, Kim RJ. Cardiovascular magnetic resonance in patients with myocardial infarction: current and emerging applications. Journal of the American College of Cardiology 55 1 (2009), pp. 1–16.
4. Ganin Y, Ustinova E, Ajakan H, Germain P, Larochelle H, Laviolette F et al. Domain-Adversarial Training of Neural Networks. 2016.
5. Dou Q, Chen C, Ouyang C, Chen H, Heng P. Unsupervised Domain Adaptation of ConvNets for Medical Image Segmentation via Adversarial Learning. 2019, pp. 93–115.
6. Kamnitsas K, Baumgartner C, Ledig C, Newcombe VFJ, Simpson JP, Kane AD et al. Unsupervised domain adaptation in brain lesion segmentation with adversarial networks. 2016.
7. Zhang L, Pereanez M, Piechnik S, Neubauer S, Petersen S, Frangi A. Multi-Input and Dataset-Invariant Adversarial Learning (MDAL) for Left and Right-Ventricular Coverage Estimation in Cardiac MRI. 2018.
8. Chen C, Ouyang C, Tarroni G, Schlemper J, Qiu H, Bai W et al. Unsupervised Multi-modal Style Transfer for Cardiac MR Segmentation. Statistical Atlases and Computational Models of the Heart. Multi-Sequence CMR Segmentation, CRT-EPiggy and LV Full Quantification Challenges. Springer International Publishing, 2020, pp. 209–219.
9. Jiang J, Hu Y, Tyagi N, Zhang P, Rimner A, Mageras G et al. Tumor-Aware, Adversarial Domain Adaptation from CT to MRI for Lung Cancer Segmentation. Vol. 11071. 2018, pp. 777–785.
10. Zhuang X. Multivariate Mixture Model for Cardiac Segmentation from Multi-Sequence MRI. 2016, pp. 581–588.

11. Stacke K, Eilertsen G, Unger J, Lundström C. A Closer Look at Domain Shift for Deep Learning in Histopathology. 2019.
12. Luo Y, Liu P, Guan T, Yu J, Yang Y. Adversarial Style Mining for One-Shot Unsupervised Domain Adaptation. 2020.
13. Chandran P, Zoss G, Gotardo P, Gross M, Bradley D. Adaptive Convolutions for Structure-Aware Style Transfer. 2021.
14. Huang X, Belongie SJ. Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization. CoRR abs/1703.06868 (2017).
15. Jafari M, Auer D, Francis S, Garibaldi J, Chen X. DRU-Net: An Efficient Deep Convolutional Neural Network for Medical Image Segmentation. 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI). 2020, pp. 1144–1148.