

# Semi-supervised Pathology Segmentation with Disentangled Representations

Haochuan Jiang<sup>1</sup>✉, Agisilaos Chartsias<sup>1</sup>, Xinheng Zhang<sup>2,3</sup>, Giorgos Papanastasiou<sup>4</sup>, Scott Semple<sup>5</sup>, Mark Dweck<sup>5</sup>, David Semple<sup>5</sup>, Rohan Dharmakumar<sup>3,6</sup>, and Sotirios A. Tsaftaris<sup>1,7</sup>

<sup>1</sup> School of Engineering, University of Edinburgh, U.K., haochuan.jiang@ed.ac.uk

<sup>2</sup> Department of Bioengineering, University of California, U.S.

<sup>3</sup> Biomedical Imaging Research Institute, Cedars-Sinai Medical Center, U.S.

<sup>4</sup> School of Computer Science and Electronic Engineering, University of Essex, U.K.

<sup>5</sup> Center for Cardiovascular Science, University of Edinburgh, U.K.

<sup>6</sup> Department of Medicine, University of California, U.S.

<sup>7</sup> The Alan Turing Institute, U.K.

**Abstract.** Automated pathology segmentation remains a valuable diagnostic tool in clinical practice. However, collecting training data is challenging. Semi-supervised approaches by combining labelled and unlabelled data can offer a solution to data scarcity. An approach to semi-supervised learning relies on reconstruction objectives (as self-supervision objectives) that learns in a joint fashion suitable representations for the task. Here, we propose Anatomy-Pathology Disentanglement Network (APD-Net), a pathology segmentation model that attempts to learn jointly for the first time: disentanglement of anatomy, modality, and pathology. The model is trained in a semi-supervised fashion with new reconstruction losses directly aiming to improve pathology segmentation with limited annotations. In addition, a joint optimization strategy is proposed to fully take advantage of the available annotations. We evaluate our methods with two private cardiac infarction segmentation datasets with LGE-MRI scans. APD-Net can perform pathology segmentation with few annotations, maintain performance with different amounts of supervision, and outperform related deep learning methods.

**Keywords:** pathology segmentation · disentangled representations · semi-supervised learning

## 1 Introduction

Deep learning models for automated segmentation of pathological regions from medical images can provide valuable assistance to clinicians. However, such models require a considerable amount of annotated data to train, which may not be easy to obtain. Pathology annotation (as opposed to anatomical), relies also on carefully detecting normal tissue areas for direct comparison. It is therefore appealing to train pathology segmentors by combining the available annotated data with larger numbers of unlabeled images in a semi-supervised learning scheme.

A typical strategy to segment pathology is to first locate the affected anatomy, e.g. the myocardium for cardiac infarction [9], and use the detected anatomy to guide the pathology prediction. However, since anatomical annotations are not always available, recent methods cascade two networks: one segments the anatomy of interest, and the second one segments the pathology [10,11,14].

Although these methods achieve accurate segmentation, they are typically fully supervised and sensitive annotations numbers, as seen in our experimental results specified in Sec. 4. Semi-supervised learning is promising to solve the issue by engaging unlabelled images. Recently, disentangled representations, i.e. structured latent spaces that are shared between labeled and unlabeled data, have provided a solution to semi-supervised learning [3,6]. These methods typically use specialized encoders to separate anatomy (a spatial tensor) and imaging information (a vector encoding image appearance) in medical image applications [3], while they involve unlabeled data through reconstruction losses.

In this paper, inspired by [3], we propose the Anatomy-Pathology Disentanglement Network (APD-Net). APD-Net constructs a space of anatomy (a spatial tensor), modality (a vector), and pathology (a spatial tensor) latent factors. Pathology is obtained by an encoder, which segments the pathology conditioned on both the image and the predicted anatomy mask. We focus on segmenting myocardial infarct, a challenging task due to its size, irregular shape and random location. APD-Net is optimized with several objectives. Among others, we introduce a novel ratio-based triplet loss to encourage the reconstruction of the pathology region by taking advantage of the pathology factor. In addition, the use of reconstruction losses, that are made possible with disentangled representations [3], makes APD-Net suitable for semi-supervised learning. Finally, we train with both predicted and real anatomy and pathology masks (in a *Teacher-Forcing* strategy [19]) to further improve performance. Our major **contributions** are summarized as follows:

- We propose a method for disentangled representations of anatomy, modality, and pathology;
- The disentanglement is encouraged with a novel ratio-based triplet loss;
- We also proposed the *Teacher-Forcing* training regime combining different scenarios of real and predicted inputs to make full use of available annotations during optimization;
- APD-Net improves the Dice score of state-of-the-art benchmarks on two private datasets for cardiac infarction segmentation when limited supervision is present, whilst maintaining performance in the full annotation setting.

## 2 Related work

**Pathology Segmentation:** A classical approach to segment pathology is by cascading organ segmentation before segmenting the pathological region. These two segmentors can be trained separately [10], or jointly [11,14]. In contrast to anatomy, pathology is small in sizes and irregular in shapes; thus, shape priors cannot be used. To train models when masks are small, the Tversky loss [15,14]

or similarly the focal loss in [1] have been proposed. Our proposed model also cascades two segmentations, by using the initial anatomy prediction to guide the subsequent pathology segmentation. However, we achieve this using disentangled representations to enable semi-supervised learning.

**Disentangled representations.** The idea of disentangled representations is to decompose the latent space into domain-invariant spatial content latent factors (known as anatomy in medical imaging) and domain-related vector style ones (here referred to as modality) [6,12]. In medical image analysis, images are disentangled in anatomical and imaging factors for the purpose of semi-supervised segmentation [3], image registration [12], and classification [18]. Image reconstruction for semi-supervised segmentation was also investigated in [5] with a simpler disentanglement of the foreground (predicted anatomy masks) and the remaining background. However, less effort has been placed on disentangling pathology. Some pioneering studies were conducted in [20], treating brain lesion segmentations as a pathology factor to synthesise pseudo-healthy images. In this paper, we also adopt a segmentation of pathology as a latent factor and combine it with disentangled anatomy and modality [3], which enables the image reconstruction task for semi-supervised learning. While we are inspired by others [20] who consider anatomy and pathology factors independently. This work is the first to learn them in a joint fashion.

### 3 Methodology

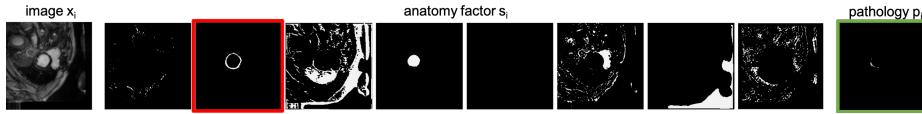
This section presents the APD-Net model. We first introduce relevant notations. Then, we specify disentanglement properties (Sec. 3.1), detail the model architecture (Sec. 3.2), and finally present the learning objectives (Sec. 3.3) with the joint training strategy depending on the different input scenarios (Sec. 3.4).

**Notation:** Let  $X, Y_{ana}, Y_{pat}$  be sets of volume slices, and the associated anatomy and pathology masks, respectively. Let  $i$  be a sample. We assume a fully labeled pathology subset  $\{x^i, y_{ana}^i, y_{pat}^i\}$ , where  $x^i \in X \subset \mathbb{R}^{H \times W \times 1}$ ,  $y_{ana}^i \in Y_{ana} := \{0, 1\}^{H \times W \times N}$ , and  $y_{pat}^i \in Y_{pat} := \{0, 1\}^{H \times W \times K}$ .  $N$  and  $K$  denote the number of anatomy, and pathology masks respectively.  $H$  and  $W$  are the image height and width. When  $Y_{pat} = \emptyset$ , it degrades to an unlabeled pathology set.<sup>8</sup> Both anatomy and pathology sets are involved in a semi-supervised fashion to segment pathology. This is achieved by learning a mapping function  $f$  that estimates anatomy and pathology given an image  $x^i$ , i.e.  $\{\hat{y}_{ana}^i, \hat{y}_{pat}^i\} = f(x^i)$ .

#### 3.1 Pathology Disentanglement

The main idea of APD-Net is to take an input image and decompose it into latent factors that relate to anatomy, pathology and image appearance (modality). This will allow inference of anatomical and pathology segmentations, whereas

<sup>8</sup> We only consider unlabeled pathology, assuming anatomy masks are available during training. Partial anatomy annotation is out of the scope of this paper.



**Fig. 1.** Visualising the disentanglement of the spatial (and binary) anatomy and pathology factors for a LGE-MRI slice.

the disentanglement of the image appearance will enable image reconstruction that is critical to enable training with unlabeled images and semi-supervised learning. Herein we consider  $C = 8$  channels of anatomy factors and  $n_z = 8$  for modality factors as in [3], and  $K = 1$  (myocardial infarct).  $s^i$  and  $p^i$  are obtained by softmax and sigmoid output activations respectively. They are then binarised (per-channel) by  $s^i - > [s^i + 0.5]$  and  $p^i - > [p^i + 0.5]$ , such that each pixel corresponds to exactly one channel. This binarisation encourages the produced anatomy factor to be modality-invariant. Finally, as in [3] gradients are bypassed in the backward pass to enable back-propagation.

Fig. 1 illustrates predicted anatomy and pathology factors for a cardiac infarct example. We make an intuitive distinction between the two factors in that the former only refers to healthy anatomical regions. Therefore, there is an overlap between the pathology factor and one or more anatomy channels. In this example (Fig. 1), the pathology (infarct in the green box) is spatially correlated with the myocardial channel (red box). Encoding pathology in the anatomy factor (i.e. entanglement of these two) is prevented, both through architecture design and with relevant losses. This will be detailed in the following sections.

### 3.2 APD-Net Architecture

APD-Net, depicted in Fig. 2, adopts modules from SD-Net [3] including anatomy  $Enc_{ana}$  and modality  $Enc_{mod}$  encoders, anatomy segmentor  $Seg_{ana}$ , and decoder  $Dec$ . They give  $s^i$ ,  $z^i$ , the anatomy mask  $\hat{y}_{ana}^i$ , and the reconstructed image  $\hat{x}^i$ .

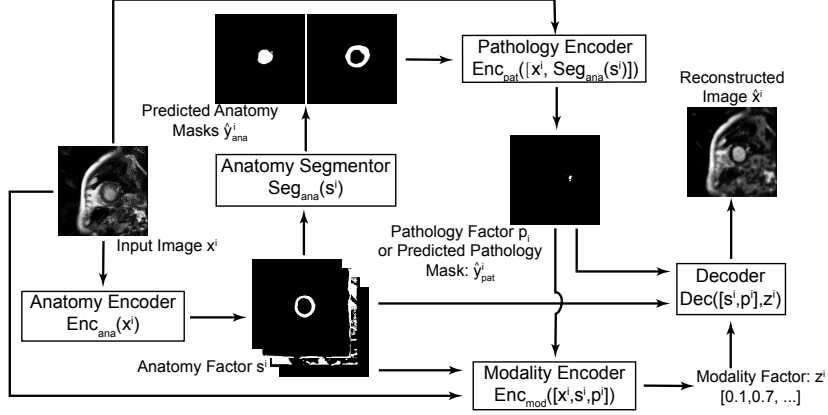
We introduce a pathology encoder  $Enc_{pat}$  following the U-Net [13] architecture. Given channel-wise concatenated  $x^i$  and  $\hat{y}_{ana}^i$ ,  $Enc_{pat}$  produces  $p^i$ .<sup>9</sup> Thus, APD-Net structurally resembles the cascaded segmentation scheme [11], enabling  $Enc_{pat}$  to focus on specific regions to locate the pathological tissue.

Finally, the image decoder receives the concatenation of  $s^i$ ,  $p^i$ , and  $z^i$  to reconstruct image  $\hat{x}^i$ , enabling unsupervised training. With nulled pathology  $p_0^i = 0$  (all elements are zero), a pseudo-healthy image ( $\hat{x}_0^i$ ) is obtained [20].

### 3.3 Individual Training Losses

APD-Net is jointly trained with losses including new supervised, unsupervised, and objectives selected from [3,4] for the task of pathology segmentation.

<sup>9</sup> Note that  $p^i$  is the same as  $\hat{y}_{pat}^i$ , i.e. the predicted pathology mask. We use  $p^i$  for disentanglement and image reconstruction, and  $\hat{y}_{pat}^i$  for pathology segmentation.



**Fig. 2.** Schematic of APD-Net. An image is encoded to anatomy factors using  $Enc_{ana}$ , and segmented with  $Seg_{ana}$  to produce anatomical segmentation masks (in this case the myocardium and left ventricle). Combined with the input, the anatomy segmentation is used to segment the pathology with  $Enc_{pat}$ . Finally, given the anatomy, the pathology, and the modality factors from  $Enc_{Mod}$ , the decoder reconstructs the input.

**Pathology Supervised Losses:** Pathology manifests in various shapes. Thus, using a mask discriminator as a shape prior [2] is not advised. In addition, pathology covers a small portion of the image leading to class imbalance between the foreground and background. To address these shortcomings, inspired by [14,17], we combine Tversky<sup>10</sup> [15] and focal loss [8]. The Tversky loss is defined as follows:  $\ell_{patT} = (\hat{y}_{pat}^i \odot y_{pat}^i) / [\hat{y}_{pat}^i + y_{pat}^i + (1 - \beta) \cdot (\hat{y}_{pat}^i - \hat{y}_{pat}^i \odot y_{pat}^i) + \beta \cdot (y_{pat}^i - \hat{y}_{pat}^i \odot y_{pat}^i)]$ , where  $\odot$  represents the element-wise multiplication. The focal loss is defined as  $\ell_{patF} = \sum_{H,W} [-y_{pat}^i (1 - \hat{y}_{pat}^i)^\gamma \log(\hat{y}_{pat}^i)]$ .

**Pathology-Masked Image Reconstruction Loss:** Typically image reconstruction is achieved by minimising the  $\ell_1$  or  $\ell_2$  loss between  $x^i$  and  $\hat{x}^i$ . However, due to the size imbalance between pathological and healthy regions, conventional full image reconstruction may ignore (average out) the pathology region. A simple, but effective solution, is to measure reconstruction performance on the pathology region by using the real pathology mask. This is implemented by a masked reconstruction  $\ell_1$  loss:  $\ell_M = \frac{1}{H \times W \times C} \sum_{H,W,C} (\frac{\lambda_{pat}}{\lambda_{ana}} \cdot y_{pat}^i + \mathbb{1}) \cdot \|x^i - \hat{x}^i\|_1$ .  $\mathbb{1}$  denotes a matrix with  $y_{pat}^i$  dimensions, where all elements are ones.<sup>11</sup>

**Ratio-based Triplet Loss:** However, this may not be adequate for accurate pathology reconstruction. We therefore penalise the model when pathology is possibly ignored by adopting a contrastive Triplet loss [16]. This is defined as  $\max(m + d_{pos} - d_{neg}, 0)$ , and minimizes the inter-class distance ( $d_{pos}$ ) compared to the intra-class ( $d_{neg}$ ) in deep feature space based on a margin ( $m$ ). We gen-

<sup>10</sup> The Dice loss can be seen as a special case of the Tversky loss [15] when  $\beta = 0.5$ .

<sup>11</sup> The  $\mathbb{1}$  matrix is added to ensure that no zero elements are multiplied with  $\|x^i - \hat{x}^i\|_1$ . Also, if  $\lambda_{pat} = 1$ , the loss reduces to the  $\ell_1$  loss.

erate *pseudo-healthy* images as negative examples,  $\hat{x}_0^i = Dec(s^i, p_0^i, z^i)$ , obtained by nulling the pathology factor, i.e.  $p_0^i$ . The deep features are calculated as a new output attached to the penultimate layer of a reconstruction discriminator (inherent from the SD-Net [4] for adversarial loss on  $\hat{x}^i$ ) for the decoder output (denoted as  $T$ ). By choosing  $T(\hat{x}^i)$  as the anchor [16], positive and negative distances are calculated as  $d_{pos} = \|T(\hat{x}^i) - T(x^i)\|_2^2$  and  $d_{neg} = \|T(\hat{x}^i) - T(\hat{x}_0^i)\|_2^2$  respectively with corresponding samples  $T(x^i)$  and  $T(\hat{x}_0^i)$ .

In practice, choosing a proper  $m$  value is challenging [21], particularly when the difference between the positive and the negative samples only lies in the small pathology region, e.g. in the cardiac infarction of Fig. 1. This will lead to an extremely small difference. Instead of optimizing the absolute margin  $m$ , we propose to alternatively minimize the Ratio-based triple loss (RT) defined as  $\ell_{RT} = \max(r + \frac{d_{pos}}{d_{neg}} - 1, 0)$ . The hyper-parameter  $r$  represents the relative margin that the positive should be closer to the anchor than the negative.

**SD-Net Losses:** We adopt optimization objectives from SD-Net [3] and its multi-modal extension [4] to the proposed framework. Specifically, the anatomy supervision Dice loss, the modality factor KL divergence, and the factor reconstruction loss are inherent from [3], while the adversarial loss on  $\hat{x}^i$  is brought from [4]. We refer SD-Net losses as  $\ell_{SDNet}$ .

### 3.4 Joint Optimization with *Teacher-Forcing* Training Strategy

A critical issue in cascaded architectures is that initial segmentation errors propagate and directly affect the second prediction. This should be taken into account particularly during training. We thus adopt the *Teaching-Forcing* strategy [19], originally applied in RNNs by engaging real rather than predicted labels.

In APD-Net, this strategy is applied on pathology segmentation  $\hat{y}_{pat}^i$ , modality factor estimation  $z^i$ , and reconstruction  $\hat{x}_i$ , which depend on real or predicted anatomy and pathology segmentations. Specifically,  $\hat{y}_{pat}^i$  can be estimated using predicted anatomy masks,  $\hat{y}_{pat}^i = Enc_{pat}([x^i, Seg_{ana}(Enc_{ana}(x^i))])$ , or real masks  $\hat{y}_{pat}^i = Enc_{pat}([x^i, y_{ana}^i])$ . Subsequently, the modality factor can be produced by the predicted pathology mask,  $z^i = Enc_{mod}([x^i, Enc_{ana}(x^i), \hat{y}_{pat}^i])$ , or the real pathology mask,  $z^i = Enc_{mod}([x^i, Enc_{ana}(x^i), y_{pat}^i])$ . Finally, real or predicted pathology contributes to reconstruction,  $\hat{x}^i = Dec([Enc_{ana}(x^i), y_{pat}^i], z^i)$  and  $\hat{x}^i = Dec([Enc_{ana}(x^i), \hat{y}_{pat}^i], z^i)$ , respectively.

We term the losses that involve the predicted, real anatomy mask, and real pathology mask as  $\ell^{PA}$ ,  $\ell^{RA}$ , and  $\ell^{RP}$  respectively. We redefine each loss as a weighted sum  $\ell = \lambda^{PA}\ell^{PA} + \lambda^{RA}\ell^{RA} + \lambda^{RP}\ell^{RP}$ , where  $\lambda^{PA}$ ,  $\lambda^{RA}$ , and  $\lambda^{RP}$  are relative weights, and  $\ell \in \{\ell_z, \ell_{RT}, \ell_{adv}, \ell_M, \ell_{KL}\}$ . Finally, the full objective is given by  $\ell_{APD-Net} = \lambda_{patT}\ell_{patT} + \lambda_{patF}\ell_{patF} + \lambda_{RT}\ell_{RT} + \lambda_M\ell_M + \ell_{SDNet}$ .

## 4 Experiments

We evaluate APD-Net on pathology segmentation using the Dice score. Experimental setup, datasets, benchmarks, and training details will be detailed below.

**Table 1.** Performance evaluation for *Data1* and *Data2*. We report test Dice scores (with standard deviation in subscript calculated by summarizing all the involved volumes) on infarct segmentation with varying infarct supervision (% infarct).

| Dataset - % infarct | <i>U-Net (unmasked)</i> | <i>Cascaded U-Net</i> | APD-Net              | <i>U-Net (masked)</i> |
|---------------------|-------------------------|-----------------------|----------------------|-----------------------|
| <i>Data1</i> -13%   | 5.4 <sub>8.1</sub>      | 36.2 <sub>17.9</sub>  | 45.3 <sub>14.4</sub> | 57.3 <sub>13.5</sub>  |
| <i>Data1</i> -25%   | 6.5 <sub>8.4</sub>      | 39.4 <sub>15.6</sub>  | 46.4 <sub>12.8</sub> | 57.3 <sub>13.2</sub>  |
| <i>Data1</i> -50%   | 26.2 <sub>13.8</sub>    | 37.1 <sub>13.8</sub>  | 46.7 <sub>11.5</sub> | 66.4 <sub>9.0</sub>   |
| <i>Data1</i> -100%  | 34.6 <sub>15.3</sub>    | 36.4 <sub>17.9</sub>  | 47.4 <sub>17.0</sub> | 65.6 <sub>10.3</sub>  |
| <i>Data2</i> -13%   | 11.5 <sub>24.4</sub>    | 25.6 <sub>23.9</sub>  | 40.0 <sub>27.0</sub> | 21.5 <sub>25.6</sub>  |
| <i>Data2</i> -25%   | 36.9 <sub>29.7</sub>    | 35.8 <sub>23.4</sub>  | 40.5 <sub>26.7</sub> | 46.7 <sub>25.6</sub>  |
| <i>Data2</i> -50%   | 15.0 <sub>14.5</sub>    | 34.4 <sub>24.2</sub>  | 38.4 <sub>17.0</sub> | 49.8 <sub>26.4</sub>  |
| <i>Data2</i> -100%  | 16.5 <sub>16.2</sub>    | 33.4 <sub>16.4</sub>  | 38.9 <sub>15.9</sub> | 45.7 <sub>28.1</sub>  |

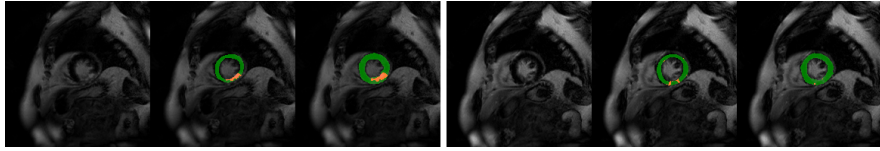
**Data:** We use two private cardiac LGE datasets acquired at the Biomedical Imaging Research Institute of the Cedars-Sinai Medical Center (*Data1*) and the Center for Cardiovascular Science of the University of Edinburgh (*Data2*), which have been approved by data ethics committees of the respective providers. Both datasets contain annotations of the myocardium and myocardial infarct. *Data1* involves 45 subjects (36 used for training) and  $224 \times 224$  dimension. *Data2* consists of 26 (mixed healthy and pathology) subjects (20 used for training), and  $192 \times 192$  dimension.

**Benchmarks:** We compare APD-Net with three benchmarks on infarct segmentation. *U-Net (masked)*: A U-Net trained on images  $x^i$  masked by the ground truth myocardium mask  $y_{ama}^i$  [9]. Masking here facilitates training, reducing the task to finding only infarcted myocardial pixels. *U-Net (unmasked)*: The U-Net is trained on images  $x_i$  without masking. This is more challenging since now the U-Net implicitly has to find infarct pixels from the whole image. *Cascaded U-Net* [11]: this trains one U-Net to segment the myocardium (with 100% supervision) and another to segment the infarct after masking the input image with the predicted myocardium (varying the number of available annotations).<sup>12</sup>

**Training details:** Training penalties are set to:  $\lambda_{patT}=1$ ,  $\lambda_{patF}=1.5$ ,  $\lambda_{RT}=1$ ,  $\lambda_M=3$ . Weights for different optimization scenarios are:  $\lambda^{PA}=1$ ,  $\lambda^{RA}=0.7$ , and  $\lambda^{RP}=0.5$ . The relative margin for  $\ell_{RT}$  is  $r=0.3$ . Other hyper-parameters include  $\beta=0.7$  in  $\ell_{patT}$ ,  $\gamma=2$  in  $\ell_{patF}$ , and the SD-Net weights defined in [3]. Due to the small data size, we do not specify validation sets. All models are trained for fixed 100 epochs, and results reported below contain averaged Dice scores and standard deviation on test data of two different splits.<sup>13</sup>

<sup>12</sup> *U-Net (masked / unmasked)* and *Cascaded U-Net* are optimized with full supervision using Tversky and focal losses, and penalized as defined in the **Training details**. In reality, *U-Net (masked)* is not a good choice since manual myocardial annotations are not always available at inference time.

<sup>13</sup> Code will be available at <https://github.com/falconjhc/APD-Net> shortly.



**Fig. 3.** Segmentation examples from *Data2*. The two panels show a good and failure infarct segmentation case. For each sample, the left image shows the input, and the next two overlay real and predicted myocardium and infarct respectively.

#### 4.1 Results and Discussion

**Semi-supervised Pathology Segmentation:** We evaluate the APD-Net performance in a semi-supervised experiment by altering the pathology supervision percentage, as seen in Table 1 respectively for the two datasets. For clarity we omit anatomy segmentation results, which are approximately 78% for both Cascaded U-Net and the proposed APD-Net for *Data1* and 64% for *Data2*.

Infarct segmentation is a challenging task, and thus all results of *Data1* and *Data2* present relatively high standard deviation, in agreement with previous literature [7]. In *Data1*, APD-Net consistently improves the Dice score of infarct prediction for all amounts of supervision, compared with both the *Cascaded U-Net* and the *U-Net (unmasked)*. Furthermore, the performance of APD-Net on small amounts of pathology labels is equivalent to the fully supervised setting.

Segmenting pathology in *Data2* is harder, as evidenced by the lower mean and higher standard deviation obtained from all methods. This could be due to the supervised methods overfitting to the smaller dataset size. APD-Net, however, overcomes this issue with semi-supervised training, and outperforms the *U-Net (masked)* at 13% annotations. While at 100% annotations, APD-Net achieves the equivalent Dice as 13% but reduces the standard deviation. More importantly, APD-Net outperforms the *Cascaded U-Net* in all setups demonstrating the benefit of image reconstruction (see also ablation studies later). Examples of correct and unsuccessful segmentations from APD-Net can be seen in Fig. 3, where the existence of sparsely-distributed annotations (right panel of Fig. 3) negatively affects supervised training.

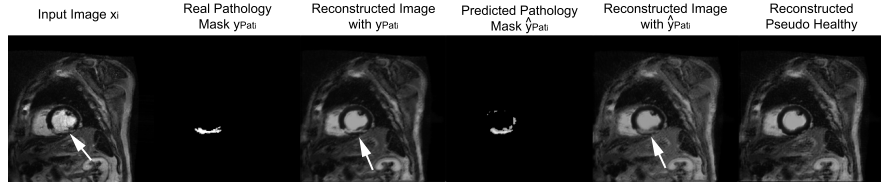
**Ablation Studies:** We evaluate the effects of critical components including the pathology-masked image reconstruction, disentanglement, teacher-forcing, and the ratio-based triplet loss with 13% and 100% infarct annotations on *Data1*. To evaluate disentanglement, we remove the modality encoder, and allow the anatomy factor to encode continuous values,  $s^i \in [0, 1]^{H \times W \times C}$ . The results presented in Table 2 show that canceling any of the ablated components hurts segmentation (except for the masked reconstruction, which at 100% performs as the proposed APD-Net). In particular, reducing annotations at 13% further decreases performance of the ablated models.

Fig. 4 depicts the effects the disentangled pathology factor on the reconstructed image. Arrows indicate the infarct region, evident when using either the real or predicted pathology to reconstruct the image. In contrast, the infarct



**Table 2.** Ablation studies on *Data1* on two infarct annotation levels (% infarct): no mask reconstruction ( $\lambda_M = 0$ ); no disentanglement (w.o. Disent.); no teacher-forcing strategy ( $\lambda^{RA} = \lambda^{RP} = 0$ ); and no ratio-based triplet loss ( $\lambda_{RT} = 0$ ).

| % annotations | $\lambda_M = 0$      | w.o. Disent.         | $\lambda^{RA} = \lambda^{RP} = 0$ | $\lambda_{RT} = 0$   | APD-Net (proposed)   |
|---------------|----------------------|----------------------|-----------------------------------|----------------------|----------------------|
| 13%           | 41.4 <sub>15.2</sub> | 14.9 <sub>8.3</sub>  | 40.7 <sub>12.4</sub>              | 38.8 <sub>14.9</sub> | 45.3 <sub>14.4</sub> |
| 100%          | 47.7 <sub>15.9</sub> | 18.4 <sub>16.6</sub> | 44.5 <sub>12.9</sub>              | 40.3 <sub>10.7</sub> | 47.4 <sub>17.0</sub> |



**Fig. 4.** Reconstruction visualizations with real, predicted pathology masks, and pseudo-healthy. White arrows point at infarct regions.

is missing when the pathology factor is nulled  $p_0^i$ , producing a pseudo-healthy image. Qualitatively, the synthetic image of the proposed APD-Net is similar to the one presented in [3]. The difference between the reconstructed and pseudo-healthy images is driven by the pathology factor. It is enhanced by the ratio based triplet loss that is essential for the desired pathology disentanglement.

## 5 Conclusions and Future Work

In this paper, we proposed the Anatomy Pathology Disentanglement Network (APD-Net) disentangling the latent space into anatomy, modality, and pathology factors. Trained in a semi-supervised scenario with reconstruction losses enabled by disentangled representation learning and joint optimization losses, APD-Net is capable of segmenting the pathology region effectively when partial pathology annotations are available. APD-Net has shown promising results in pathology segmentation using partial annotations and improved performance compared to other related baselines in the literature.

However, APD-Net still follows the cascaded pathology segmentation strategy that can propagate errors from the first to the second segmentation, which is not solved fundamentally even with the *Teacher-Forcing*. Diseases that deform the anatomical structure (e.g., brain tumour) cannot be predicted by cascading. In addition, we only tested the proposed APD-Net in myocardial infarct, where the pathology manifests as high-intensity regions within the myocardium. As future work, we aim to explore direct pathology segmentation methods without predicting the relevant anatomy. Meanwhile, we plan to investigate extensions of the current APD-Net that are more general and do not restrict to myocardial infarct, while also engaging multi-modal images that would offer complementary anatomical information. Finally, we will test our method on public datasets with more examples to further validate 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. Abraham, N., Khan, N.M.: A novel focal tvsky loss function with improved attention u-net for lesion segmentation. In: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019). pp. 683–687. IEEE (2019)
2. Chartsias, A., Joyce, T., Papanastasiou, G., Semple, S., Williams, M., Newby, D., Dharmakumar, R., Tsaftaris, S.A.: Factorised spatial representation learning: Application in semi-supervised myocardial segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 490–498. Springer (2018)
3. Chartsias, A., Joyce, T., Papanastasiou, G., Semple, S., Williams, M., Newby, D.E., Dharmakumar, R., Tsaftaris, S.A.: Disentangled representation learning in cardiac image analysis. *Medical image analysis* **58**, 101535 (2019)
4. 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)
5. Dey, R., Hong, Y.: Compnet: Complementary segmentation network for brain mri extraction. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 628–636. Springer (2018)
6. Huang, X., Liu, M.Y., Belongie, S., Kautz, J.: Multimodal unsupervised image-to-image translation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 172–189 (2018)
7. Karim, R., Bhagirath, P., Claus, P., Housden, R.J., Chen, Z., Karimaghloo, Z., Sohn, H.M., Rodríguez, L.L., Vera, S., Albà, X., et al.: Evaluation of state-of-the-art segmentation algorithms for left ventricle infarct from late gadolinium enhancement mr images. *Medical image analysis* **30**, 95–107 (2016)
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. Moccia, S., Banali, R., Martini, C., Muscogiuri, G., Pontone, G., Pepi, M., Caiani, E.G.: Development and testing of a deep learning-based strategy for scar segmentation on cmr-lge images. *Magnetic Resonance Materials in Physics, Biology and Medicine* **32**(2), 187–195 (2019)
10. Morshid, A., Elsayes, K.M., Khalaf, A.M., Elmohr, M.M., Yu, J., Kaseb, A.O., Hassan, M., Mahvash, A., Wang, Z., Hazle, J.D., et al.: A machine learning model to predict hepatocellular carcinoma response to transcatheter arterial chemoembolization. *Radiology: Artificial Intelligence* **1**(5), e180021 (2019)
11. Pang, Y., Hu, D., Sun, M.: A modified scheme for liver tumor segmentation based on cascaded fcn. In: Proceedings of the International Conference on Artificial Intelligence, Information Processing and Cloud Computing. pp. 1–6 (2019)
12. Qin, C., Shi, B., Liao, R., Mansi, T., Rueckert, D., Kamen, A.: Unsupervised deformable registration for multi-modal images via disentangled representations.

- In: International Conference on Information Processing in Medical Imaging. pp. 249–261. Springer (2019)
13. 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)
  14. Roth, K., Konopczyński, T., Hesser, J.: Liver lesion segmentation with slice-wise 2d tiramisu and tversky loss function. arXiv preprint arXiv:1905.03639 (2019)
  15. 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)
  16. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 815–823 (2015)
  17. Tran, G.S., Nghiem, T.P., Nguyen, V.T., Luong, C.M., Burie, J.C.: Improving accuracy of lung nodule classification using deep learning with focal loss. *Journal of healthcare engineering* **2019** (2019)
  18. van Tulder, G., de Bruijne, M.: Learning cross-modality representations from multi-modal images. *IEEE transactions on medical imaging* **38**(2), 638–648 (2018)
  19. Williams, R.J., Zipser, D.: A learning algorithm for continually running fully recurrent neural networks. *Neural computation* **1**(2), 270–280 (1989)
  20. Xia, T., Chartsias, A., Tsafaris, S.A.: Pseudo-healthy synthesis with pathology disentanglement and adversarial learning. *Medical Image Analysis* **64**, 101719 (2020)
  21. Zakharov, S., Kehl, W., Planche, B., Hutter, A., Ilic, S.: 3d object instance recognition and pose estimation using triplet loss with dynamic margin. In: 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 552–559. IEEE (2017)