THE UNIVERSITY *of* EDINBURGH

Edinburgh Research Explorer

Temporal Consistency Objectives Regularize the Learning of Disentangled Representations

OPEN ACCESS

Temporal Consistency Objectives Regularize the Learning of Disentangled Representations

Gabriele Valvano[1,2] Agisilaos Chartsias[2], Andrea Leo[1], and Sotirios A. Tsaftaris[2]

[1] IMT School for Advanced Studies Lucca, Piazza S. Francesco, Lucca 55100 LU, Italy
[2] School of Engineering, University of Edinburgh, West Mains Rd, Edinburgh EH9 3FB, UK

Abstract. There has been an increasing focus in learning interpretable feature representations, particularly in applications such as medical image analysis that require explainability, whilst relying less on annotated data (since annotations can be tedious and costly). Here we build on recent innovations in style-content representations to learn anatomy, imaging characteristics (appearance) and temporal correlations. By introducing a self-supervised objective of predicting future cardiac phases we improve disentanglement. We propose a temporal transformer architecture that given an image conditioned on phase difference, it predicts a future frame. This forces the anatomical decomposition to be consistent with the temporal cardiac contraction in cine MRI and to have semantic meaning with less need for annotations. We demonstrate that using this regularization, we achieve competitive results and improve semi-supervised segmentation, especially when very few labelled data are available. Specifically, we show Dice increase of up to 19% and 7% compared to supervised and semi-supervised approaches respectively on the ACDC dataset. Code is available at: https://github.com/gvalvano/sdtnet.

Keywords: Disentangled Representations · Semi-supervised Learning · Cardiac Segmentation.

1 Introduction

Recent years have seen significant progress in the field of machine learning and, in particular, supervised learning. However, the success and generalization of such algorithms heavily depends learning suitable representations [2]. Unfortunately, obtaining them usually requires large quantities of labelled data, which need expertise and in many cases are expensive to obtain.

It has been argued [3] that good data representations are those separating out (disentangling) the underlying explanatory factors into disjoint subsets. As

a result, latent variables become sensitive only to changes in single generating factors, while being relatively insensitive to other changes [2]. Disentangled representations have been reported to be less sensitive to nuisance variables and to produce better generalization [16]. In the context of medical imaging, such representations offer: i) better interpretability of the extracted features; ii) better generalization on unseen data; iii) and the potential for semi-supervised learning [5]. Moreover, disentanglement allows interpretable latent code manipulation, which is desirable in a variety of applications, such as modality transfer and multi-modal registration [5, 10, 13].

Medical images typically present the spatial information about the patient's anatomy (shapes) modulated by modality-specific characteristics (appearance). The SDNet framework [5] is an attempt to decouple anatomical factors from their appearance towards more explainable representations. Building on this concept, we introduce a new architecture that drives the model to learn anatomical factors that are both spatially and temporally consistent. We propose a new model, namely: Spatial Decomposition and Transformation Network (SDTNet).

The main **contributions** of this paper are: **(1)** we introduce a modality invariant transformer that, conditioned on the temporal information, predicts future anatomical factors from the current ones; **(2)** we show that the transformer provides a self-supervised signal useful to improve the generalization capabilities of the model; **(3)** we achieve state of the art performance compared to SDNet for semi-supervised segmentation at several proportions of labelled data available; **(4)** and show for the first time preliminary results of cardiac temporal synthesis.

2   Related Works

2.1   Learning good representations with temporal conditioning

The world surrounding us is typically affected by smooth temporal variations and is known that temporal consistency plays a key role for the development of invariant representations in biological vision [17]. However, despite that temporal correlations have been used to learn/propagate segmentations in medical imaging [1, 12], their use as a learning signal to improve representations remains unexplored. To the best of our knowledge, this is the first work to use spatiotemporal dynamics to improve disentangled representations in cardiac imaging.

Outside the medical imaging community, we find some commonalities of our work with Hsieh et al. [7], who address the challenge of video frame prediction decomposing a video representation in a time-invariant *content* vector and a time-dependent *pose* vector. Assuming that the content vector is fixed for all frames, the network aims to learn the dynamics of the low-dimensional pose vector. The predicted pose vector can be decoded together with the fixed content features to generate a future video frame in pixel space. Similarly, we decompose the features space in a fixed and a time-dependent subset (*modality* and *anatomy*). However, our objective is not merely predicting a future temporal frame, but we use the temporal prediction as a self-supervised signal to ameliorate the quality

of the representation: i.e. we constrain its temporal transformation to be smooth. By doing so, we demonstrate that we can consistently improve the segmentation capabilities of the considered baselines.

2.2   Spatial Decomposition Network (SDNet)

Here, we briefly review a recent approach for learning disentangled anatomy-modality representations in cardiac imaging, upon which we build our model.

The SDNet [5] can be seen as an autoencoder taking as input a 2D image $x \sim X$ and decomposing it into its anatomical components $s = f_A(x)$ and modality components $z = f_M(x)$. The vector z is modelled as a probability distribution $Q(z|X)$ that is encouraged to follow a multivariate Gaussian, as in the VAE framework [9]. s is a multi-channel output composed of binary discrete maps. A decoder $g(\cdot)$ uses both s and z to reconstruct the input image $\tilde{x} = g(s, z) \approx x$. An additional network $h(\cdot)$ is supervisedly trained to extract the heart segmentation $\tilde{y} = h(s)$ from s, while an adversarial signal forces $\tilde{y}$ to be realistic even when few pairs of labelled data are available, enabling semi-supervised learning.

While SDNet was shown to achieve impressive results in semi-supervised learning, it still requires human annotations to learn to decouple the cardiac anatomy from other anatomical factors. Furthermore, it doesn't take advantage of any temporal information to learn better anatomical factors: as a result they are not guaranteed to be temporally correlated.

3   Proposed Approach

Herein, we address the above limitations, by a simple hypothesis: components s of different cardiac phases should be similar within the same cardiac cycle and their differences, if any, should be consistent across different subjects. To achieve this we introduce a new neural network $T(\cdot)$ in the SDNet framework that, conditioned on temporal information, regularizes the anatomical factors such that they can be consistent (e.g. have smooth transformations) across time. Obtaining better representations will ultimately allow improved performance in the segmentation task, too. $T(\cdot)$ is a modality-invariant transformer that 'warps' the s factors learnt by the SDNet according to the cardiac phase. Furthermore, by combining the current z factors with the predicted s factors for future time points, one can reconstruct the future frames in a cardiac sequence: e.g., given time $t_1 < t_2$, we have $\tilde{x}_{t_2} = g(T(s_{t_1}), z_{t_1}) \approx x_{t_2}$. Our model is shown in Figure 1. Below we focus our discussion on the design of the transformer and the training costs, all other network architectures follow that of SDNet [5]. In the following, t, dt are scalars, while remaining variables are considered as tensors.

3.1   Spatial Decomposition and Transformation Network (SDTNet)

The transformer $T(\cdot)$ takes as input the binary anatomical factors s (Figure 2) and their associated temporal information t. Under the assumption that the
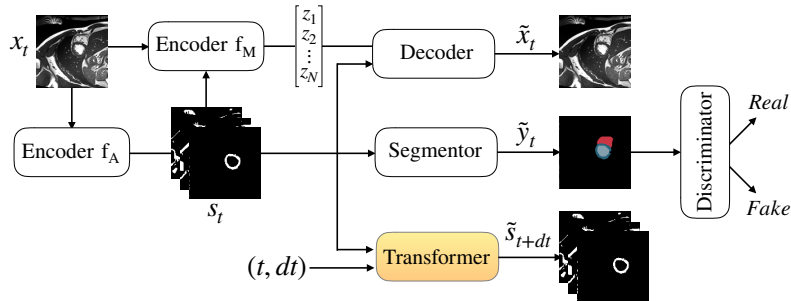
Fig. 1: SDTNet block diagram. The transformer (in yellow) predicts the future anatomical factors conditioned on the temporal information. The future frame can be generated by the decoder using $\tilde{s}_{t+dt}$ and the current z factor.

Fig. 2: Anatomical factors extracted by the SDTNet from the image on the left.

modality factors remain constant throughout the temporal dimension (e.g. the heart contracting from extra-diastole to extra-systole), the transformer must deform the current anatomy s_t such that, given a temporal change dt, it estimates s_{t+dt}, ie. the anatomy of image x_{t+dt} when given as input. Using this prediction $\tilde{s}_{t+dt} = T(s_t, t, dt)$ together with the fixed modality factors z_t, we should be able to correctly reconstruct the image at the future time point $\tilde{x}_{t+dt}$. By capturing the temporal dynamics of the anatomical factors, the transformer guides their generation to be temporally coherent, resulting in a self-supervised training signal, that is the prediction error of future anatomical factors.

3.2   Transformer design

After testing several architecture designs for the transformer, we found that the best results could be obtained by adapting a UNet [14] to work with binary input/output conditioned on temporal information on the bottleneck.

Temporal information, the tuple (t, dt), is encoded via an MLP consisting of 3 fully connected layers, arranged as 128-128-4096, with the output reshaped to $16 \times 16 \times 16$. This information is concatenated at the bottleneck of the UNet where features maps have resolution $16 \times 16 \times 64$, to condition the transformer and control the required deformation. To encourage the use of the temporal features and retain the notion of the binary inputs, the features at the bottleneck and of the MLP are bounded in $[0, 1]$, using a sigmoid activation function.

We hypothesised that it would be easier to model differential changes to anatomy factors. Thus, we added a long residual connection between the UNet

input and its output. We motivate this by observing that the anatomical structure that mostly changes in time is the heart: thus learning the spatial transformation should be similar to learning to segment the cardiac structure in the binary tensor s: a task that the UNet is known to be effective at solving. The output of the transformer is binarized again (key for disentanglement), as in [5].

3.3   Cost Function and Training

The overall cost function is the following weighted sum:

$$Loss = \lambda_0 \cdot L_S + \lambda_1 \cdot L_{US} + \lambda_2 \cdot L_{ADV} + \lambda_3 \cdot L_{TR} \ , \tag{1}$$

where $\lambda_0 = 10$, $\lambda_1 = 1$ and $\lambda_2 = 10$ as in [5], and $\lambda_3 = 1$ found experimentally. $\boldsymbol{L_S}$ is the cost associated to the supervised task (segmentation) and can be written as $L_S = L_{DICE}(y, \tilde{y}) + 0.1 \cdot L_{CE}(y, \tilde{y})$, where y and $\tilde{y}$ are the ground truth and predicted segmentation masks, respectively; L_{DICE} is the differentiable Dice loss evaluated on left ventricle, right ventricle and myocardium, while L_{CE} is the weighted cross-entropy on these three classes plus the background (with class weights inversely proportional to the number of pixels for the class).

$\boldsymbol{L_{US}}$ is the cost associated to the unsupervised task and can be decomposed as $L_{US} = |\tilde{x} - x| + \lambda_{KL} \cdot D_{KL}[Q(z|X)||N(0, I)] - MI(\tilde{x}, z)$. The first term is the mean absolute error between the input and the reconstruction, while the second term is the KL divergence between $Q(z|X)$ and a Normal Gaussian (with λ_{KL}=0.1). The last term is the mutual information between the reconstruction $\tilde{x}$ and the latent code z and is approximated by an additional neural network, as in the InfoGAN framework [6]. By maximizing the mutual information between the reconstruction and z, we prevented posterior collapse and constrained the decoder $g(\cdot)$ to effectively use the modality factors.

$\boldsymbol{L_{ADV}}$ is the adversarial loss of a Least-Squares GAN [11], used to discriminate ground truth from predicted segmentations in the unsupervised setting.

$\boldsymbol{L_{TR}}$ is the loss associated to the self-supervised signal, computed as the differentiable Dice loss between $\tilde{s}_{t+dt}$ and s_{t+dt}. This error serves as a proxy for the reconstruction error of future cardiac phases $|x_{t+dt} - g(T(s_t), z_t)|$. In practice, we find it much easier to train $T(\cdot)$ with a loss defined in the anatomy space rather than one on the final reconstruction: in fact, the gradients used to update the network parameters can flow into $T(\cdot)$ directly from its output layer, rather than from that of the decoder $g(\cdot)$.

The model was optimized using the Exponential Moving Average (EMA): we maintained a moving average of the parameters during training, and employed their average for testing. The learning rate was scheduled to follow a triangular wave [15] in the range 10^{-4} to 10^{-5} with a period of 20 epochs. Both EMA and the learning rate scheduling facilitated comparisons, allowing to detect wider and more generalizable minima (hence, reducing loss fluctuations). We used Adam [8] with an early stopping criterion on the segmentation loss of a validation set.
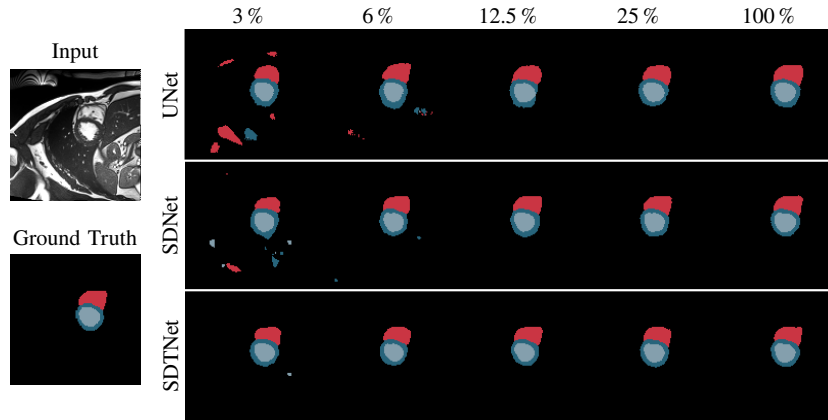
Fig. 3: Comparison of predicted segmentations obtained from the UNet, SDNet, SDTNet after being trained with different percentages of the labelled data.

4   Experiments and Discussion

4.1   Data and Preprocessing

Data. We used ACDC data from the 2017 Automatic Cardiac Diagnosis Challenge [4]. These are 2-dimensional cine-MR images acquired using 1.5T and 3T MR scanners from 100 patients, for which manual segmentations for the left ventricular cavity (LV), the myocardium (MYO) and the right ventricle (RV) are provided in correspondence to the end-systolic (ES) and end-diastolic (ED) cardiac phases. ES and ED phase instants are also provided. We used a 3-fold cross validation and randomly divided the data to obtain 70 MRI scans for training, 15 for validation and 15 for the test sets.

Preprocessing. After removing outliers outside 5th and 95th percentiles of the pixel values, we removed the median and normalized the images on the interquartile range, centering each volume approximately around zero.

Training. Since our objective was to introduce temporal consistency in the anatomical factors rather then predicting the whole cardiac cycle, we split the cine MRI sequences in two halves: i) temporal frames in the ED-ES interval; ii) temporal frames from ES to the end of the cardiac cycle. The latter frames were reversed in their temporal order, to mimic once again the cardiac contraction: as a result, we avoided the inherent uncertainty associated to the transformations of frames in the middle of the cardiac cycle. Finally, we applied data augmentation at run-time, consisting of rotations, translations and scaling of each 2D slice.

4.2   Results

Semi-supervised segmentation We compared SDTNet to the fully supervised training of a UNet and to the semi-supervised training of SDNet in a segmentation task, varying the percentage of labelled training samples. As Figure 3

Labels	UNet	SDNet	SDTNet	Improvement
100%	80.03 ±0.38	85.11 ±0.73	85.83 ±0.40	0.72
25%	77.55 ±1.02	81.64 ±0.96	83.69 ±0.37	2.05*
12.5%	71.04 ±1.71	78.07 ±1.52	79.48 ±0.82	1.41*
6%	59.20 ±1.38	72.18 ±1.91	74.22 ±0.57	2.04*
3%	44.89 ±9.52	56.89 ±2.48	63.74 ±1.59	6.85*

Table 1: DICE scores comparing SDTNet and other baselines at various proportions of available labeled data. The last column shows the average improvement of SDTNet over SDNet. Asterisks denote statistical significance ($p < 0.01$).

Time 0.0   0.33   0.66   1.0 $\tilde{x}_{t=1} - \tilde{x}_{t=0}$
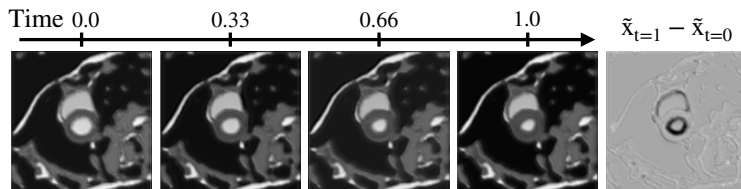

Fig. 4: Interpolation on the temporal axis between ED and ES phases. The images are obtained by fixing the modality-dependent factors $z_{t=0}$ and using the anatomical factors $\tilde{s}_{t>0}$ predicted for future time points. In Acrobat, clicking on the rightmost image animates frames showing the predicted cardiac contraction.

and Table 1 show, the SDTNet consistently outperforms the others, especially at lower percentages of labelled pairs in the training set. Furthermore, SDTNet exhibits lower variance in its predictions, so it's more consistent. A paired Wilcoxon test demonstrated most of these improvements to be statistically significant. We find that the transformer forces the anatomical decomposition to follow more "semantic" disentanglement even with little human annotations. This translates to better segmentation results. While secondary to the thesis of the paper, both the SDNet and the SDTNet outperform the UNet.

Cardiac synthesis Figure 4 shows that it is possible to predict future cardiac phases from ED through ES by using the predicted anatomical factors $\tilde{s}_{t>0}$ together with the modality factors $z_{t=0}$. We note that this is the first attempt of deep learning-based temporal synthesis in cardiac albeit preliminary. Note that we train the transformer with both pathological and healthy subjects and it thus predicts average temporal transformations. Conditioning also with prior pathology information and validation of synthesis are left as future work.

5 Conclusion

We introduced a self-supervised objective for learning disentangled anatomy-modality representations in cardiac imaging. By leveraging the temporal infor-

mation contained in cine MRI, we introduced a spatiotemporal model in SD-Net [5], improving its generalization capabilities in the semi-supervised setting at several proportions of labelled data available. Also, the resulting approach considerably outperforms the fully-supervised baseline, confirming the potential for semi-supervised and self-supervised training in medical imaging.

References

1. Bai, W., Suzuki, H., Qin, C., Tarroni, G., Oktay, O., Matthews, P.M., Rueckert, D.: Recurrent neural networks for aortic image sequence segmentation with sparse annotations. In: MICCAI. pp. 586–594. Springer (2018)
2. Bengio, Y., Courville, A., Vincent, P.: Representation learning: A review and new perspectives. IEEE PAMI **35**(8), 1798–1828 (2013)
3. Bengio, Y., et al.: Learning deep architectures for AI. Foundations and trends in Machine Learning **2**(1), 1–127 (2009)
4. Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.A., Cetin, I., Lekadir, K., Camara, O., Ballester, M.A.G., et al.: Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved? IEEE TMI **37**(11), 2514–2525 (2018)
5. 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)
6. Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., Abbeel, P.: Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In: NeurIPS. pp. 2172–2180 (2016)
7. Hsieh, J.T., Liu, B., Huang, D.A., Fei-Fei, L.F., Niebles, J.C.: Learning to decompose and disentangle representations for video prediction. In: NeurIPS. pp. 517–526 (2018)
8. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. ICLR (2015)
9. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. ICLR (2014)
10. Lee, H.Y., Tseng, H.Y., Huang, J.B., Singh, M., Yang, M.H.: Diverse image-to-image translation via disentangled representations. In: ECCV. pp. 35–51 (2018)
11. Mao, X., Li, Q., Xie, H., Lau, R.Y.K., Wang, Z., Smolley, S.P.: On the effectiveness of least squares generative adversarial networks. IEEE PAMI (2018)
12. Qin, C., Bai, W., Schlemper, J., Petersen, S.E., Piechnik, S.K., Neubauer, S., Rueckert, D.: Joint learning of motion estimation and segmentation for cardiac mr image sequences. In: MICCAI. pp. 472–480. Springer (2018)
13. Qin, C., Shi, B., Liao, R., Mansi, T., Rueckert, D., Kamen, A.: Unsupervised deformable registration for multi-modal images via disentangled representations. arXiv preprint arXiv:1903.09331 (2019)

14. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI. pp. 234–241. Springer (2015)
15. Smith, L.N.: Cyclical learning rates for training neural networks. In: 2017 IEEE WACV. pp. 464–472. IEEE (2017)
16. Van Steenkiste, S., Locatello, F., Schmidhuber, J., Bachem, O.: Are disentangled representations helpful for abstract visual reasoning? arXiv preprint arXiv:1905.12506 (2019)
17. Wood, J.N.: A smoothness constraint on the development of object recognition. Cognition **153**, 140–145 (2016)