Factorised Spatial Representation Learning: Application in Semi-supervised Myocardial Segmentation

Agisilaos Chartsias[1(✉)], Thomas Joyce[1], Giorgos Papanastasiou[2],
Scott Semple[2], Michelle Williams[2], David Newby[2], Rohan Dharmakumar[3],
and Sotirios A. Tsaftaris[1]

[1] Institute for Digital Communications, School of Engineering,
University of Edinburgh, West Mains Rd, Edinburgh EH9 3FB, UK
`agis.chartsias@ed.ac.uk`
[2] Edinburgh Imaging Facility QMRI, Centre for Cardiovascular Science,
Edinburgh EH16 4TJ, UK
[3] Cedars Sinai Medical Center, Los Angeles, CA, USA

Abstract. The success and generalisation of deep learning algorithms heavily depend on learning good feature representations. In medical imaging this entails representing anatomical information, as well as properties related to the specific imaging setting. Anatomical information is required to perform further analysis, whereas imaging information is key to disentangle scanner variability and potential artefacts. The ability to factorise these would allow for training algorithms only on the relevant information according to the task. To date, such factorisation has not been attempted. In this paper, we propose a methodology of latent space factorisation relying on the cycle-consistency principle. As an example application, we consider cardiac MR segmentation, where we separate information related to the myocardium from other features related to imaging and surrounding substructures. We demonstrate the proposed method's utility in a semi-supervised setting: we use very few labelled images together with many unlabelled images to train a myocardium segmentation neural network. Specifically, we achieve comparable performance to fully supervised networks using a fraction of labelled images in experiments on ACDC and a dataset from Edinburgh Imaging Facility QMRI. Code will be available at https://github.com/agis85/spatial_factorisation.

1 Introduction

The effectiveness of any (deep or shallow) learning algorithm lies in learning good feature representations. These must be maximally informative for the task at hand, whilst being invariant to unrelated information (e.g. variations in imaging, noise, etc), so that they can generalise to unseen examples [5]. Invariance to some factors, e.g. translations, can be attributed to the architecture, for instance with the use of convolution and max-pooling, but invariance to more complex factors

is achieved by the learning process, and specifically encouraged by regularisers (explicit regularisation) or data augmentation (implicit regularisation) [1].

At a high level the aim is to keep relevant but discard irrelevant information, however which information is relevant is strongly task dependent. In this paper we are interested in the related task of decomposing the input into meaningful components (or factors), which offers many benefits. Critically it enables preserving factors not directly relevant to the primary task, which may otherwise be discarded when driven by pure supervised learning. It is then possible to reuse parts of a factorised representation for related tasks or for transfer learning to other domains. Further, by capturing specific properties of the data, such representations become easier to interpret, an aspect of currently heated debate in deep learning with dedicated workshops on the topic (e.g. http://interpretable.ml). Finally, finding (and preserving) the factors of variation is *de facto* necessary for generative models, in order to be able to (re)produce realistic results.

Factorised representations are a recent topic in deep learning [6,7,9,11,14,15]. These works focus on decomposing feature representations into discrete or continuous latent vectors. At present, there has not been any work on learning factorised representations that include spatial components, which are of particular interest for spatially equivariant tasks using fully convolutional networks (such as segmentation and registration).[1] Here we propose a spatial decomposition network (SDNet), that decomposes input images into a spatial map containing anatomical information and a latent vector of image intensity information (and residual anatomical information), leveraging the cycle-consistency loss [21], originally proposed for style transfer. Specifically, we train two networks: one that learns a decomposition into spatial and non-spatial latent factors, and one that learns to reconstruct the input image using the decomposed representation. We demonstrate our method in semi-supervised myocardium segmentation, using a small amount of labelled but a large pool of unlabelled cardiac cine MR images. In this application, our method learns to decompose the shape and location of the myocardium from information related to surrounding structures and pixel intensities (related to scanner properties and other imaging characteristics).

In summary, our contributions are the following: (a) We propose a new method for disentangling images into a spatial map and a continuous vector, which is directly applicable to medical images for representing anatomical and non-anatomical information. (b) We show properties of the decomposed latent space by generating examples using latent space arithmetic. (c) We demonstrate the utility of our method in a semi-supervised myocardium segmentation task, where the learned high-level topological knowledge allows the network to retain performance in a low data regime.

---

[1] Concurrent work with ours, introduced auxiliary variables and combined them with a spatial representation for the task of image translation [2,10].

2   Related Work

Learning Factorised Representations: To date interest has centred on representing factors of variation as independent latent variables, using Autoencoders [7] or Variational Autoencoders (VAE) [15] to decompose classification related factors from remaining image reconstruction factors. VAE were used for unsupervised learning of factorised representations, where the factors of variation are discovered throughout the learning process [9,11]. A generative model combining VAE with Generative Adversarial Networks (GAN) was proposed in [14] to decompose the input into image classes and remaining factors. Further, InfoGAN was proposed in [6], in which mutual information between a latent variable and the generated images is maximised. More recently, feature decompositions were proposed for video data to separate foreground from background [19], and motion from content [18]. These methods learn decomposed representations in terms of continuous or discrete variables; however, spatial information could be directly represented in a convolutional map, and this would be useful when the learning task is semantic segmentation. Our proposed method produces a decomposition as a combination of spatial and non-spatial information. This makes our learned representation directly applicable to segmentation tasks.

Semi-supervised Segmentation: Using unlabelled data to guide learning is appealing and has been exploited by the community. In [3] an iterative method was proposed, where a CNN is alternately trained on labelled and post-processed unlabelled sets. GANs were used in [20], for a gland segmentation task, involving supervised and unsupervised adversarial costs. Another approach [4] aims to minimise the distance between embeddings of labelled and unlabelled examples by comparing them in feature space. Semi-supervised learning with GANs was also proposed for semantic segmentation. The discriminator classifies between real and synthetic segmentation masks produced by the generator in [12], while in [17] the generator is used to increase the dataset size and the discriminator performs segmentation. Our method differs from these in that we introduce both adversarial and cycle losses to push mask generation to be spatially aligned with the image and avoid the need for post-processing as in [3]. Also we do not require any pairs of image and masks for discriminator training as in [20], and we retain all information, in contrast to [4] which preserves only task relevant information.
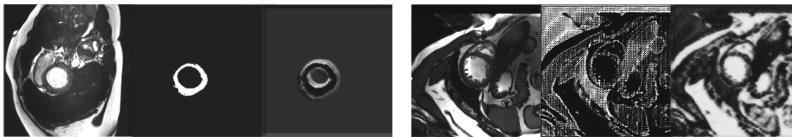

Fig. 1. Input images, segmentation masks and reconstructions produced by a Cycle-GAN. Left: high weight on segmentation, right: high weight on reconstruction.

3   Proposed Approach: The SDNet

Motivation: A useful latent representation is one that describes the data well. Spatial (segmentation) maps can be considered a form of latent variable that allows visual inspection of what a network learns. At the same time, an easy (unsupervised) way to see whether a latent representation captures the data is to use a decoder to reconstruct the input. In fact, even CycleGANs are autoencoders: they encode (and decode) the input via an intermediate output and thus inspire the design of our approach. Yet they have problems particularly when the intermediate output is discretised (a binary mask) and supervised losses are introduced. Their performance heavily depends on the weighting of the losses, as shown in Fig. 1. If the segmentation loss is weighted higher than the reconstruction loss, it is not possible to reconstruct the input since the binary mask does not contain enough information for the transformation. When differently weighted, information is stored in the binary mask ruining semantics. This confirms findings of others, that a CycleGAN resolves the many-to-one/one-to-many problem by storing low-frequency information in the output image [8]. We can see that the two losses are antagonistic, and a standard CycleGAN is not suitable as is. We need to introduce variables that break the many-to-one problem, encouraging a balance between the losses to achieve good segmentation and reconstruction.
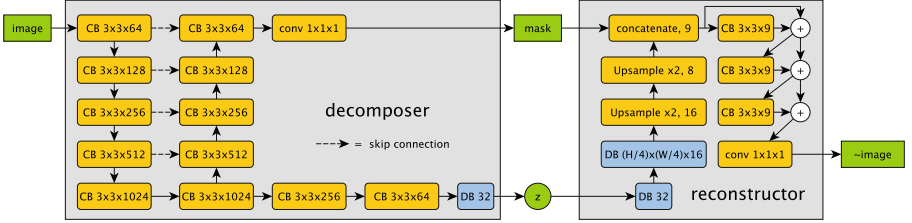

Fig. 2. Schematic of SDNet: an image is decomposed as a spatial representation of anatomy (in our case myocardial mask M) and a latent vector Z that captures other anatomical and imaging characteristics. Both mask and Z are used to reconstruct the input. The model consists of several convolutional (CB) and dense blocks (DB). BatchNormalization and LeakyRelu activations are used throughout.

SDNet: Our model is comprised of two interconnected neural networks, a "decomposer" and a "reconstructor", as illustrated in Fig. 2. The former decomposes an input 2D image (slice in a cine acquisition) into two components: a spatial representation of the myocardium in the form of a binary mask, and a latent representation of the remaining anatomical and imaging features in the form of a vector. Thus, the mask is an image having pixel to pixel correspondences with the input and is inherently spatial, whereas the other representation is a vector representing information in a high level way that is not directly spatial. The reconstructor receives the two representations and aims to synthesise

the original input image. Given a successful decomposition, the binary mask acts
as a guide defining where the reconstructed myocardium should be. The role of
the latent feature variable is then to learn some topology around the myocardium
and fill the necessary intensity patterns, and allow for many-to-many mappings.

Costs: More formally, let f and g be the decomposer and reconstructor. Given
an image slice X_i, we aim to learn weights of f to decompose into a mask M and
a 16 dimensional vector Z, that is $f(X_i) = \{f_M(X_i), f_Z(X_i)\} = \{M, Z\}$, and the
weights of g to remap the decomposition back to an image $g(f_M(X_i), f_Z(X_i))$.

In a semi-supervised setup data comes from a labelled set $S_L =$
$\{X_i, M_i\}_{i \in [1,N]}$ and an unlabelled set $S_U = \{X_j\}_{j \in [1,M]}$ where usually $M > N$.
We now define the following losses. Firstly, a reconstruction loss from autoen-
coding an image, $L_{rec}(f, g) = \mathbb{E}_X[\|X - g(f(X))\|_1]$. Secondly, two super-
vised losses when having images with corresponding masks M_X, $L_M(f) =$
$\mathbb{E}_X[Dice(M_X, f_M(X))]$, and $L_I(f, g) = \mathbb{E}_X[\|X - g(M_X, f_Z(X)))\|_1]$. Finally,
an adversarial loss using an image discriminator D_X, as $A_I(f, g, D_X) =$
$\mathbb{E}_X[D_X(g(f(X)))^2 + (D_X(X) - 1)^2]$. Networks f and g are trained to max-
imise this objective against an adversarial discriminator trained to minimise
it. Similarly, we define an adversarial loss using a mask discriminator D_M as
$A_M(f) = \mathbb{E}_{X,M}[D_M(f_M(X))^2 + (D_M(M) - 1)^2]$. Both adversarial losses are
based on [13]. The overall cost function is defined as:

$$\lambda_1 L_M(f) + \lambda_2 A_M(f, D_M) + \lambda_3 L_{rec}(f, g) + \lambda_4 L_I(f, g) + \lambda_5 A_I(f, g, D_X)$$

The loss for images from the unlabelled set does not contain the first and fourth
terms. The λ are experimentally set to 10, 10, 1, 10 and 1 respectively.

Implementation Details: The decomposer follows a U-Net [16] architecture
(see Fig. 2), and its last layer outputs a segmentation mask of the myocardium
via a sigmoid activation function. The model's deep spatial maps contain down-
sampled image information, which is used to derive the latent vector Z through
a series of convolutions and fully connected layers, with the final output being
passed through a sigmoid so Z is bounded. Following this, an architecture with
three residual blocks is employed as the reconstructor (see Fig. 2).

The spatial and continuous representations are not explicitly made indepen-
dent, so during training the model could still store all information needed for
reconstructing the input as low values in the spatial mask, since finding a map-
ping from a spatial representation to an image is easier than combining two
sources of information, namely the mask and Z. To prevent this, we apply a
step function (i.e. a threshold) at the spatial input of the reconstructor to bina-
rise the mask in the forward pass. We store the original values and bypass the
step function during back-propagation, and apply the updates to the original
non-binary mask. Note that the binarisation of the mask only takes place at the
input of the reconstructor network and is not used by the discriminator.

4   Experiments and Discussion

4.1   Data and Baselines

ACDC: We use data from the 2017 ACDC Challenge[2] containing cine-MR images from patients with various disease. Images were acquired in 1.5T or 3T MR scanners, with resolution between 1.22 and $1.68\,\mathrm{mm}^2$/pixel and the number of phases varying between 28 to 40 images per patient. We resample all volumes to $1.37\,\mathrm{mm}^2$/pixel resolution and normalise in the range $[-1, 1]$.

QMRI: We also use cine-MR data acquired at Edinburgh Imaging Facility QMRI with a 3T scanner, of 28 healthy patients, each having a volume of 30 frames. The spatial resolution is $1.406\,\mathrm{mm}^2$/pixels with a slice thickness 6 mm, matrix size 256×216 and field of view 360 mm $\times$ 303.75 mm.

Baselines: We use as a *fully-supervised baseline* a standard U-Net network trained with a Dice loss, similar to most participants of the ACDC challenge. We also consider a *semi-supervised baseline*, shorthanded as GAN below, by adding a GAN loss to the supervised loss to allow adversarial training [12].
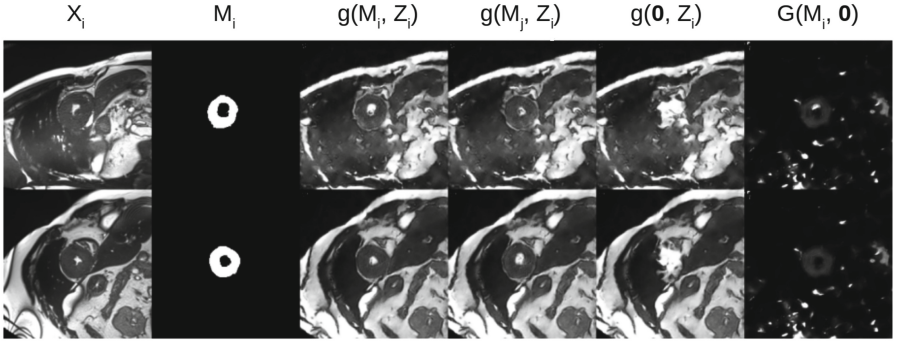

Fig. 3. Reconstructions using different M_i and Z_i combinations (see text for details).

4.2   Latent Space Arithmetic

As a demonstration of our learned representation, in Fig. 3 we show reconstructions of input images from the training set using different combinations of masks and Z components. In the first three columns, we show the original input with the predicted mask and the input's reconstruction. Next, we take the spatial representation M_j from one image and combine it with the Z_i component of the other image, and vice versa. As shown in the figure (4th column) the intensities and the anatomy around the myocardium remains unchanged, but the myocardial shape and position, which are encoded in the mask, change to that of the

---

[2] https://www.creatis.insa-lyon.fr/Challenge/acdc/index.html.

second image. The final two columns show reconstructions using a null mask (i.e. $M_i = \mathbf{0}$) and the correct Z_i in 5th column, or using the original mask with a $Z_i = \mathbf{0}$ in 6th column. In the first case, the produced image does not contain myocardium, whereas in the second case the image contains only myocardium and no other anatomical or MR characteristics.
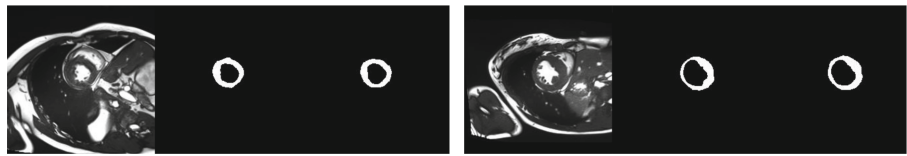

Fig. 4. Two examples of segmentation performance: input, prediction and ground truth.

4.3 Semi-supervised Results

The utility of a factorised representation becomes evident in semi-supervised learning. Qualitatively in Fig. 4 we can see that our method closely follows ground truth segmentation masks (example from ACDC held-out test set).

To assess our performance quantitatively we train a variety of setups varying the number of labelled training images whilst keeping the unlabelled fixed (in both ACDC and QMRI cases). We train SDNet and the baselines (U-Net and GAN), test on held-out test sets, and use 3-fold cross validation (with 70%, 15%, 15% of the volumes used in training, validation and test splits respectively). Results are shown in Table 1. For reference a U-Net trained with supervision on the full ACDC and QMRI datasets achieves a Dice score of 0.817 and 0.686 respectively. We can see that even when the number of labelled images is very low, our method is able to achieve segmentation accuracy considerably higher than the other two methods. As the number of labelled images increases, all models achieve similar accuracy.

Table 1. Myocardium Dice scores on ACDC and QMRI data. For training, 1200 unlabelled and varying numbers of labelled images were used. Masks for adversarial training came from the dataset, but do not correspond to any training images.

	ACDC					QMRI			
Labelled images	284	142	68	34	11	157	78	39	19
U-Net	0.782	0.657	0.581	0.356	0.026	0.686	0.681	0.441	0.368
GAN	**0.787**	0.727	0.648	0.365	0.080	**0.795**	0.756	0.580	0.061
SDNet	0.771	**0.767**	**0.731**	**0.678**	**0.415**	0.794	**0.772**	**0.686**	**0.424**

5   Conclusion

We presented a method that decomposes images into spatial and (non-spatial) latent representations employing the cycle-consistency principle. To the best of our knowledge this is the first work to investigate spatial representation factorisation, in which one factor of the representation is inherently spatial, and thus well suited to spatial tasks. We demonstrated its applicability in semi-supervised myocardial segmentation. In the low-data regime ($\approx$1% of labelled with respect to unlabelled data) it achieves remarkable results, showing the power of the proposed learned representation. We leave as future work generative extensions, where we learn statistical distributions of our embeddings (as in VAEs).

References

1. Achille, A., Soatto, S.: Emergence of invariance and disentangling in deep representations. In: ICML Workshop Principled Approaches to Deep Learning (2017)
2. Almahairi, A., Rajeswar, S., Sordoni, A., Bachman, P., Courville, A.: Augmented CycleGAN: learning many-to-many mappings from unpaired data. In: Dy, J., Krause, A. (eds.) Proceedings of the 35th International Conference on Machine Learning, Proceedings of Machine Learning Research, vol. 80, pp. 195–204. PMLR, Stockholm (2018). http://proceedings.mlr.press/v80/almahairi18a.html
3. Bai, W., et al.: Semi-supervised learning for network-based cardiac MR image segmentation. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI 2017. LNCS, vol. 10434, pp. 253–260. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66185-8_29
4. Baur, C., Albarqouni, S., Navab, N.: Semi-supervised deep learning for fully convolutional networks. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI 2017. LNCS, vol. 10435, pp. 311–319. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66179-7_36
5. Bengio, Y., Courville, A., Vincent, P.: Representation learning: a review and new perspectives. IEEE PAMI **35**(8), 1798–1828 (2013)
6. Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., Abbeel, P.: InfoGAN: interpretable representation learning by information maximizing generative adversarial nets. In: NIPS, pp. 2172–2180 (2016)
7. Cheung, B., Livezey, J.A., Bansal, A.K., Olshausen, B.A.: Discovering hidden factors of variation in deep networks. arXiv:1412.6583 (2014)
8. Chu, C., Zhmoginov, A., Sandler, M.: CycleGAN: a Master of Steganography. arXiv:1712.02950 (2017)
9. Higgins, I., et al.: beta-VAE: Learning basic visual concepts with a constrained variational framework. In: ICLR (2017)
10. Huang, X., Liu, M.-Y., Belongie, S., Kautz, J.: Multimodal Unsupervised Image-to-image Translation. In: ECCV (2018)
11. Kim, H., Mnih, A.: Disentangling by factorising. In: ICML (2018)

12. Luc, P., Couprie, C., Chintala, S., Verbeek, J.: Semantic segmentation using adversarial networks. arXiv preprint arXiv:1611.08408 (2016)
13. Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Smolley, S.P.: On the Effectiveness of Least Squares Generative Adversarial Networks. arXiv:1712.06391 (2017)
14. Mathieu, M.F., Zhao, J.J., Zhao, J., Ramesh, A., Sprechmann, P., LeCun, Y.: Disentangling factors of variation in deep representation using adversarial training. In: NIPS, pp. 5040–5048 (2016)
15. Narayanaswamy, S., et al.: Learning disentangled representations with semi-supervised deep generative models. In: NIPS, pp. 5927–5937 (2017)
16. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
17. Souly, N., Spampinato, C., Shah, M.: Semi and weakly supervised semantic segmentation using generative adversarial network. arXiv:1703.09695 (2017)
18. Tulyakov, S., Liu, M.-Y., Yang, X., Kautz, J.: Mocogan: decomposing motion and content for video generation. arXiv:1707.04993 (2017)
19. Vondrick, C., Pirsiavash, H., Torralba, A.: Generating videos with scene dynamics. In: NIPS, pp. 613–621 (2016)
20. Zhang, Y., et al.: Deep adversarial networks for biomedical image segmentation utilizing unannotated images. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI 2017. LNCS, vol. 10435, pp. 408–416. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66179-7_47
21. Zhu, J.-Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: ICCV, pp. 2242–2251 (2017)