I. SUPPLEMENTARY MATERIAL

1) Implementation Details: In this section, we give the details about the meta-parameter choices and sampling strategies. In both the super-resolution (SR) and segmentation (Seg) models, the weight decay regularisation term is weighted by $\lambda_2 = 5 \times 10^{-6}$, and gradient descent learning-rate is fixed to $lr = 0.001$. The weight of global priors is chosen experimentally to be $\lambda_1 = 0.01$. Mini batch-size, which is the number of samples used for each back-propagated gradient update, is set to be 8 samples. The models are trained with full images without a need for patch extraction since the cardiac 2D MR image stack size is relatively smaller compared to the available GPU memory (Nvidia GTX-1080).

In the SR problem, the through-plane upsampling factor is fixed to $K = 5$ and the synthetic low-resolution training samples are generated simply by filtering high-resolution images with a Gaussian blurring kernel ($\sigma = 4.0$ mm) along the through plane direction. The blurring operation is followed by a decimation operator along the same image dimension. In the segmentation problem, the Sorensen-Dice loss was tested in the experiments as an alternative to the cross-entropy loss, yet we observed a degraded performance since the latter is a smoother function.

2) Network Structure: In this section, we give the network structures of the autoencoder (AE) and the predictor, which together build the T-L network. The details are provided in Table I and II. As can be seen in the tables, residual connections are not used in our models (10-18 layers) since they do not provide significant accuracy gains for smaller networks as reported in [1]. Additionally, non-linear layers are not applied on the lower dimensional latent representations since that would further constrain the autoencoder. The number of the hidden units (64) is chosen experimentally, and optimal configurations can be explored to improve the reconstruction performance. The input layer of the AE model is extended to multi-label segmentation maps through one-hot image representation, where each label is converted into a separate channel at the input. Lastly, in both models (AE and PR) each convolution layer, except the last one, is followed by batch-normalisation for better convergence behaviour.

TABLE I: Structure of the predictor model: The model maps the input HR intensity image (120x120x60) to the latent space and generates a 64-dim representation. The size, number, and stride of the learnt convolution (Conv) kernels are provided. The filters operate on different image scales ($S1$-$S4$) and each convolution operation is followed by a non-linear unit (ReLU).

		Size	Stride	# Kernels	Non-linearity
$S1$	Conv	(f:3,3,3)	(s:1,1,1)	(N:32)	ReLU
	Conv	(f:3,3,3)	(s:1,1,1)	(N:32)	ReLU
$S2$	Conv	(f:3,3,3)	(s:2,2,2)	(N:64)	ReLU
	Conv	(f:3,3,3)	(s:1,1,1)	(N:64)	ReLU
$S3$	Conv	(f:3,3,3)	(s:2,2,2)	(N:128)	ReLU
	Conv	(f:3,3,3)	(s:1,1,1)	(N:128)	ReLU
$S4$	Conv	(f:3,3,3)	(s:2,2,2)	(N:256)	ReLU
	Conv	(f:3,3,3)	(s:1,1,1)	(N:1)	ReLU
	FC	-	-	(N:64)	None

TABLE II: Structure of the autoencoder (AE) model: The encoder part maps the given input segmentation map (120x120x60) to the latent space through convolution (Conv) and fully-connected (FC) layers. The decoder part recovers the input from the low-dimensional representation and outputs a segmentation map (120x120x60). The size, number, and stride of the learnt convolution (Conv) kernels are provided. The filters operate on different image scales ($S1$-$S4$) and each convolution operation is followed by a non-linear unit (ReLU).

		Kernel	Stride	# Kernels	NonLin
$S1$	Conv	(f:3,3,3)	(s:2,2,1)	(N:16)	ReLU
	Conv	(f:3,3,3)	(s:1,1,1)	(N:16)	ReLU
$S2$	Conv	(f:3,3,3)	(s:2,2,2)	(N:32)	ReLU
	Conv	(f:3,3,3)	(s:1,1,1)	(N:32)	ReLU
$S3$	Conv	(f:3,3,3)	(s:2,2,2)	(N:64)	ReLU
	Conv	(f:3,3,3)	(s:1,1,1)	(N:64)	ReLU
$S4$	Conv	(f:3,3,3)	(s:3,3,3)	(N:1)	ReLU
HC	FC	-	-	(N:64)	None
	FC	-	-	(N:125)	ReLU
$S4$	Deconv	(f:7,7,7)	(s:3,3,3)	(N:64)	ReLU
	Conv	(f:3,3,3)	(s:1,1,1)	(N:64)	ReLU
$S3$	Deconv	(f:4,4,4)	(s:2,2,2)	(N:32)	ReLU
	Conv	(f:3,3,3)	(s:1,1,1)	(N:32)	ReLU
$S2$	Deconv	(f:4,4,4)	(s:2,2,2)	(N:16)	ReLU
	Conv	(f:3,3,3)	(s:1,1,1)	(N:16)	ReLU
$S1$	Deconv	(f:4,4,1)	(s:2,2,1)	(N:16)	ReLU
	Conv	(f:3,3,3)	(s:1,1,1)	(N:3)	None

REFERENCES

[1] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. *arXiv preprint arXiv:1609.04802*, 2016.