6. Supplementary

6.1. Label Generation

For the gereration of the labels that we use for extending our training dataset, we augment the existing labels as follows: a new label is composed either of one or the combination of two randomly chosen labels. One or both of them have a random orientation and x-y shift (roll) in the image. The labels can randomly be eroded or dilated, or not. The number of final metastases in a label used for generating the synthetic data varies between one and 90, in accordance with the number of metastases found in the natural images.

6.2. DICE scores with off-the-shelf segmentor

As an additional evaluation of the quality of generated metastases, we have run the segmentor used in the paper on the images generated for the test set. From the obtained values, we can observe that MetGAN obtains the best DICE score of all studied methods, and this value is in accordance with the original performance of the segmentor.

Table 5. Evaluation of MetGAN and compared methods with the off-the-shelf segmentor. We can observe that MetGAN obtains the best DICE and Recall, while SIFA obtains best precision. We speculate that the latter is due to the trivial samples SIFA generates, which are easy to guess by the segmentor. However, as it fails to translate correctly the domain, the segmentor has a high false positive rate, resulting in a low recall. Common issues of Pix2Pix and CycleGAN are randomly placed metastases, which translate to lower segmentation scores (especially for CycleGAN). SPADE and SEAN with VAE also produce samples that are easy to find (thus high precision), but incorrectly generated contrasts (see Figure 4, images have bright background and dark foreground objects) confuse the segmentor, resulting in a low recall score. Simple SPADE performs worse than with the variational autoencoder, as the burden on the network is increased, and the generated samples are less realistic. Lastly, we attribute RedGAN's low scores to the artefact-ridden backgrounds that are generated.

Network	DICE ↑	Precision ↑	Recall ↑
Pix2Pix	0.53	0.75	0.43
CycleGAN	0.10	0.12	0.12
RedGAN	0.19	0.53	0.13
SPADE	0.45	0.43	0.55
SPADE+VAE	0.61	0.82	0.51
SEAN+VAE	0.59	0.76	0.53
SIFA	0.50	**0.85**	0.45
MetGAN	**0.8**	0.8	**0.8**

6.3. Setup of SOTA Methods

All networks that we have used for comparison in our paper have been adapted for our data. All inputs are of size 256×256 pixels, and normalized in the range [-1,1] for the autofluorescence and cancer channel, and [0,1] for the label.

Pix2Pix was set up with the default settings, recieving 2 channel inputs (autofluorescence and label concatenated), and 1 channel output (cancer channel).

CycleGAN has been set up either as Pix2Pix (2 channels for Domain A - autofluorescence + label, 1 channel for Domain B - cancer channel), or with 2 times concatenated cancer channel as output. The scores in our paper are the better of the two, and were obtained with the first setup.

RedGAN was trained with the label as the main input to its SPADE network, the autofluorescence as input to its VAE. The segmentor from [20] (also used in MetGAN) was employed as its segmentor network.

SPADE+VAE and SEAN+VAE were trained with the label as main input, and the autofluorescence as input to the VAE. Simple SPADE only uses the label as input.

SIFA has been modified to take a 3 channel input: autofluorescence, label, and a zero image, and to produce a 3 channel output that concatenates the the same image of the cancer channel 3 times.

Unmentioned parameters were used with their default values as suggested by their authors.

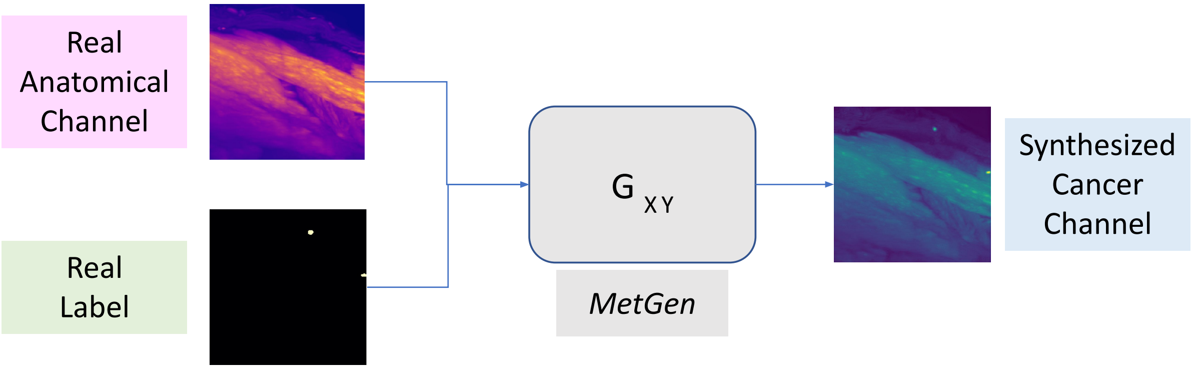

Figure 8. Generation of synthetic data. For generating new samples, we combine a real autofluorescence channel image and a generated label (method described above) as inputs into the trained generator, MetGen.

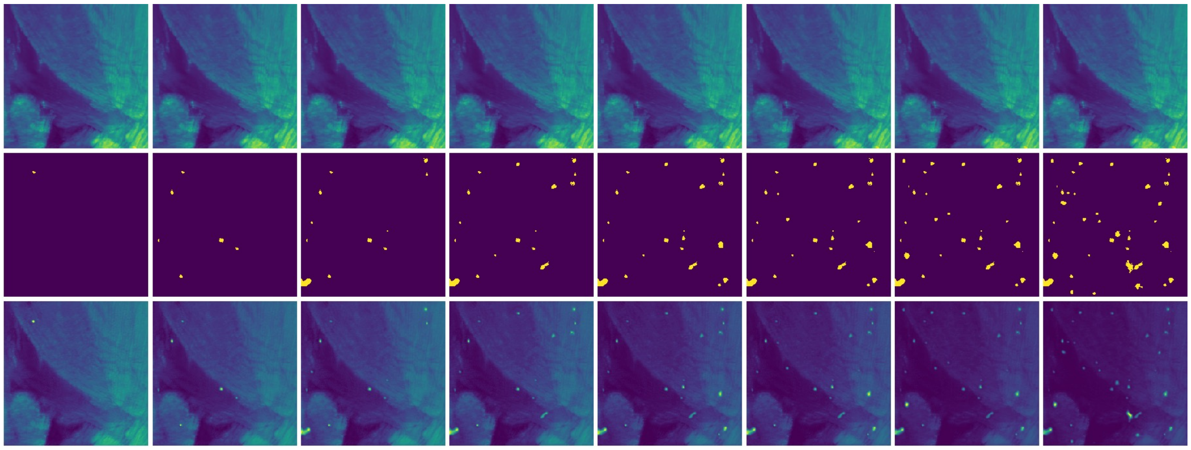

Figure 9. Generation of synthetic data with incremental amount of metastases in the label image. It can be observed that the labels directly enforce the existence of a tumor on the exact spatial location. For this label generation method, we have sampled random real metastases in 3D space from all available labels, incrementally placed them in an empty volume, and generated their projection.

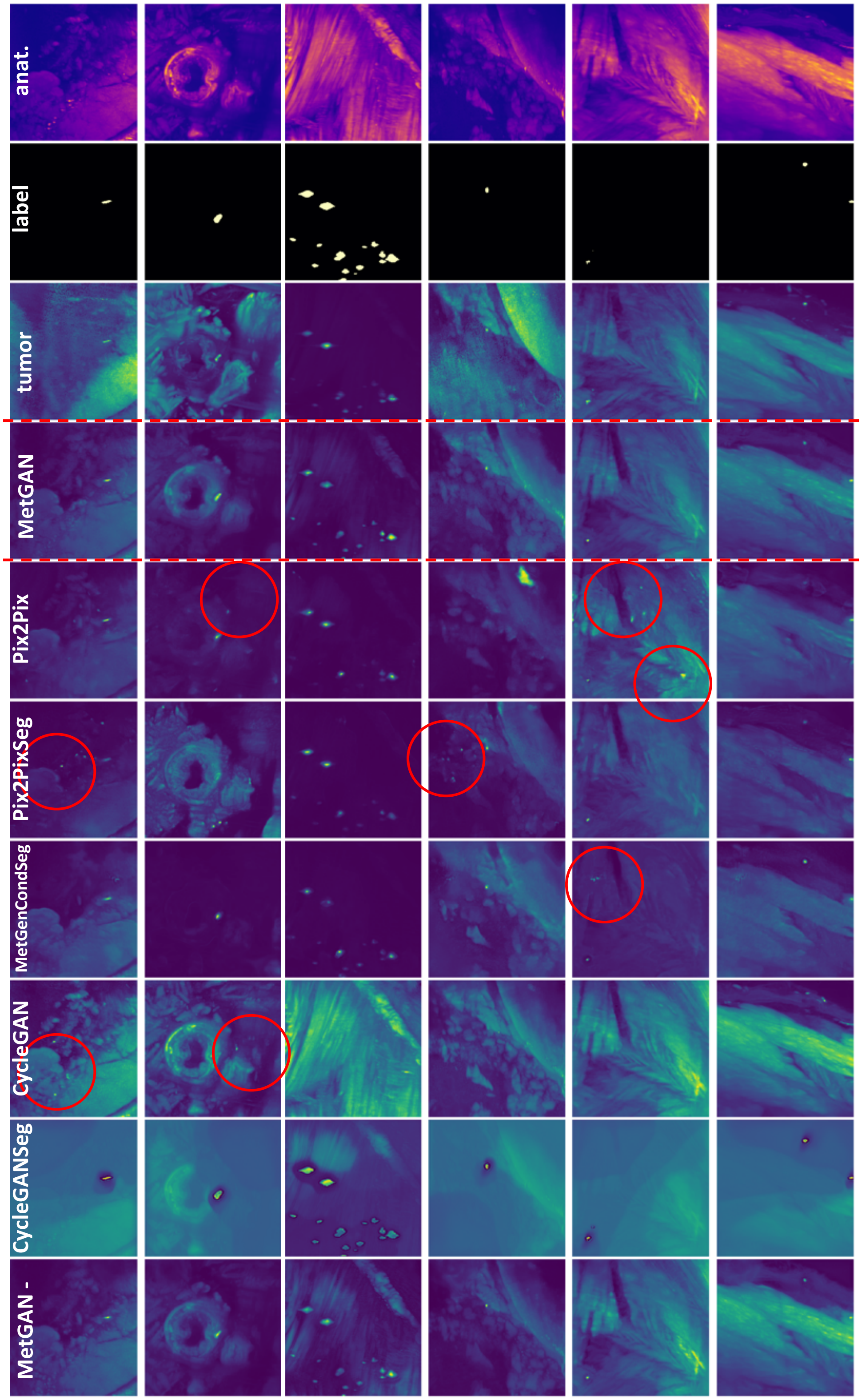

Figure 10. Qualitative results from our proposed generative method compared to baseline and ablated models: Pix2Pix, Pix2PixSeg, MetGenCondSeg, CycleGAN, CycleGANSeg, MetGAN-. We find that our method generates the most realistic looking tumour images, with tumours in the correct locations. The baseline and ablated models fails at respecting the imposed semantic map and lacks in the domain translation, either by not using features from the anatomical domain or by keeping too many.

Table 6. Comparison of a selection of SOTA applications to our method. Abbreviations: AF - autofluorescence, G - generator, D - discriminator, CD - conditional discriminator, S - segmentor, SD- segmentation discriminator, VAE - variational autoencoder.

Method	Input		Network Architecture	Summary	Main Difference to our work	Common Differences
	Label	AF				Generator Architecture **Image-to-im.:** Semantics in both domain No label input **Semantic Synthesis:** No domain adaptation
Pix2Pix [11]		Yes	G+CD	Image-to-image Translation (Paired, conditional).	No cycle consistency. No Segmentation loss. No label input.	
Adapted Pix2Pix	Yes	Yes	G+CD			
Cycle GAN [45]		Yes	2xG + 2xD	Image-to-image Translation (Unpaired).	No Segmentation loss. No label input. No pair loss.	
Adapted CycleGAN	Yes	Yes	2xG + 2xD			
SIFA [5]		Yes	2xG, 2xD, S + SD	Medical Image-to-image Translation with feature alignment and segmentation loss.	No pair loss. Segmentation loss not compared to ground truth.	
Adapted SIFA	Yes	Yes	2xG, 2xD, S + SD			
SPADE [21]	Yes	(VAE)	G+D	Semantic Synthesis in target domain Can accept style input into VAE.	Synthesis in target domain Generator has to create background.	
SEAN [46]	Yes	(VAE)	G+D	Styled Semantic Synthesis in target domain. Can accept style input into VAE.	Synthesis in target domain. Generator has to create background.	
RedGAN [22]	Yes	(VAE)	G+D+S	Medical Semantic Synthesis in target domain. Can accept style input into VAE. Constrained by segmentor.	Synthesis in target domain. Generator has to create background.	
Stanford *et al.* [24]		Yes	2xG, 2xD	3D Cycle GAN for medical image to image translation (CT).	Same as CycleGAN.	
Xu *et al.* [38]		Yes	2xG, 2xD + attention	Attention-guided tumour generation in brain MRI.	Location of tumours decided by the network (not conditional synthesis) no image translation.	
Liu *et al.* [17]		Yes	2xG, 2xD	Image-to-image translation (non contrast to contrast CT).	Same as CycleGAN.	
Zhang *et al.* [42]		Yes	2xG, 2xD, 2xS	Image-to-image transaltion with segmentation in both domains.	Same as CycleGAN.	
Abhishek *et al.* [1]	Yes		G+CD	Conditional synthesis in target domain (label to image translation).	No domain adaptation. No segmentation.	
Wu *et al.* [36]	Yes		G+CD	Conditional infilling of tumour in an image in the target domain.	No domain adaptation, operate directly on image in target domain.	
Liu *et al.* [16]	Yes	Yes	2xG + 2xD + Semantic Alignment	Image-to-image translation followed by separate semantic alignment.	Multi-step process (style transfer followed by generation of semantics).	