# METGAN: Generative Tumour Inpainting and Modality Synthesis in Light Sheet Microscopy

Izabela Horvath [1,2]
izabela.horvath@tum.de

Johannes Paetzold [1,2]
johannes.paetzold@tum.de

Oliver Schoppe [1,2]
oliver.schoppe@tum.de

Rami Al-Maskari [1,2]
rami.al-maskari@tum.de

Ivan Ezhov [2,3]
ivan.ezhov@tum.de

Suprosanna Shit [2,3]
suprosanna.shit@tum.de

Hongwei Li [2,4]
hongwei.li@tum.de

Ali Ertürk [1,5]
erturk@helmholtz-muenchen.de

Bjoern Menze [4]
bjoern.menze@uzh.ch

[1] Institute for Tissue Engineering and Regenerative Medicine, Helmholtz Zentrum Munich, Germany
[2] Department of Computer Science, Technical University of Munich, Germany
[3] TranslaTUM Center for Translational Cancer Research, Munich, Germany
[4] Department of Quantitative Biomedicine, University of Zurich, Switzerland
[5] Institute for Stroke and Dementia Research, Ludwig Maximilian University of Munich, Germany

## Abstract

*Novel multimodal imaging methods are capable of generating extensive, super high resolution datasets for preclinical research. Yet, a massive lack of annotations prevents the broad use of deep learning to analyze such data. In this paper, we introduce a novel generative method which leverages real anatomical information to generate realistic image-label pairs of tumours. We construct a dual-pathway generator, for the anatomical image and label, trained in a cycle-consistent setup, constrained by an independent, pretrained segmentor. Our method performs two concurrent tasks: domain adaptation and semantic synthesis, which, to our knowledge, has not been done before. The generated images yield significant quantitative improvement compared to existing methods that specialize in either of these tasks. To validate the quality of synthesis, we train segmentation networks on a dataset augmented with the synthetic data, substantially improving the segmentation over the baseline.*

## 1. Introduction

Recently, the combination of fluorescence microscopy and tissue clearing has enabled the generation of single-cell resolution, terabyte sized 3D datasets of whole specimens and organs [29, 33, 44]. These datasets facilitate the quantitative study of disease or age-induced anatomical alterations, as well as drug targeting, in human organs or animal models. An intriguing feature of such datasets is the multi-channel nature of the data. An autofluorescence channel, denoted as ”anatomical channel”, images general tissue, and a pathology channel marks objects of interest such as tumours (”tumour channel”) using fluorescence dyes. The sheer size and multi-channel characteristics of these data require the use of high throughput deep learning methods to analyze and segment them [31, 20, 27]. However, data volume and complexity increase the manual annotation cost, as the timeframe required for manual labeling of a single scan of a whole mouse can span up to two months. This evidently motivates the development of new approaches for synthetic image-label pair generation and data augmentation [19].

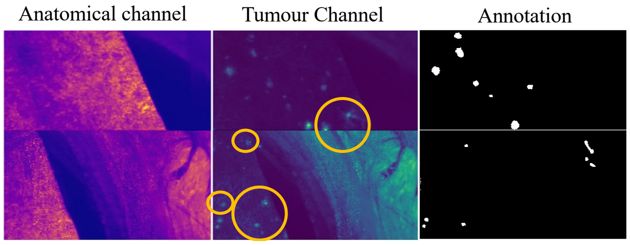

Figure 1. Motivation: the characteristic dim metastases can only be identified by multiple adjustments of contrast, and only from certain angles. Such properties lead to inconsistent annotations, even from experts. First, these inconsistencies make the labeling very expensive. Second, training segmentation networks on these datasets is only stable for large numbers of annotated samples. Both of these aspects motivate the need for our generative tumour inpainting to create labeled datasets.

Nevertheless, our experiments showed us that existing methods often fail to generate semantically correct images when underlying annotations used for training are noisy. Importantly, current approaches either perform image translation of structures that are visible in both domains, or generate them, based on labels, in the target domain. This leaves the generator with the task of synthesising backgrounds that fail to take advantage of structurally rich priors. The generated images are semantically correct, but often trivial. Therefore, these aspects motivated the technical development of *MetGAN*, a generative method that performs domain translation and semantic synthesis simultaneously, generating realistic images in the target domain that show structures imposed by the desired label.

Application wise, we present a generative adversarial network (GAN) based approach, which leverages the advantages of fluorescence microscopy datasets, namely a high SNR and multiple channels, to generate realistic data. Our application focuses on a dataset pertaining metastatic spread in a whole mouse organism. Based on information in the anatomical channel, we generate synthetic images in the pathology (cancer) domain, with objects of interest (tumours) placed in user-defined locations. Thus, we derive the generated samples from existent and distinctive priors, without the additional burden on the generator of having to synthesise both diverse backgrounds and foreground. Additionally, our method solves characteristic inconsistencies in foreground data and labels (see Figure 1).

The use of generative adversarial networks for synthetic data generation and augmentation [28, 24] to improve segmentation and classification tasks is a current topic of interest. Goodfellow *et al.* introduced the concept of GANs [8], which was successfully improved and extended towards image-to-image translation [11, 45, 6], with one of the key applications being medical imaging [40, 13, 32]. Within these approaches, a representation of a structure of interest is translated into another domain or multiple other domains such as different imaging modalities or contrasts[37, 17]. For this application, most studies use a Cycle-GAN inspired architecture (for unpaired data), or conditional GANs (such as Pix2Pix), where applications focus on aligned and paired datasets. Based on such results Cohen *et al.* showed that the use of distribution matching losses can lead to hallucinating structures (such as tumours), which translated to medical misdiagnosis [7], indicating the need for advanced image synthesis techniques; for example losses which punish the generation of unwanted elements. Thus, in order to enforce the preservation and correct translation of the semantics of the data, many authors constrain their GANs with one or more additional segmentation networks [10, 43, 5, 39]. Another approach is dividing the task into two stand-alone steps: adaptation, followed by semantic alignment [16]. This concept improved the combination of appearance and semantic adaptation.

Another common use of GANs in computer vision and medical imaging is to generate semantically guided images in a target domain [21, 46, 14, 36, 12, 2, 1]. A common feature of these applications is the use of Spade ResNet blocks in the generator, sometimes further constrained by an additional segmentor [22]. For other medical images, label-based conditional GANs have been used for data infill [15] and augmentation for underrepresented classes [18], to improve segmentation [42], as well as classification [35, 4, 25]. While GAN-based approaches have been used to generate synthetic tumor images [38], to the best of our knowledge no coupling between images and labels has been achieved. A summary of relevant state-of-the-art methods and their difference to our proposed method can be found in the Supplementary material.

***Contributions:*** 1) We develop a novel generative model, *MetGAN*, to synthesize realistic microscopic images of tumour metastases, based on real autofluorescence image information and arbitrarily placed tumour labels. Our model consists of a dual-pathway generator, *MetGen*, trained in a cycle-consistent setup, and further constrained by an independent, pretrained segmentor. Our novelty lies in: the generator architecture, the addition of a passive segmentor in the cycle-consistent training setup, and the additional constraint with a pair-wise loss for improved domain translation, as well as the task itself: concurrent medical domain translation and semantic synthesis. 2) We present qualitative results of our generated microscopic tumour images, and we quantitatively evaluate the error comparative to the real ground truth images. This shows the superiority of our model over existing state of the art methods. 3) We extensively validate our method in an ablation study and a downstream segmentation task, where we use our generated tumour images to train a segmentation network. We show that training solely on generated image-label pairs achieves identical performance as training on a large set of real data. Furthermore, augmentation of the real dataset with synthtetic data improves lesion detection.

## 2. Methods

### 2.1. Architecture

In this work, we propose a cycle-consistent 2D framework for domain translation and semantic tumour inpainting, using a customized GAN, whose architecture is depicted in Figure 3. Our setup is inspired by CycleGAN, a proven model in domain translation tasks [45].

It has been shown that networks which employ simple batch normalization layers tend to lose semantic information when it comes to label-based generation [21]. This aspect motivated us to construct proposed our generator network, *MetGen*, with an additional pathway, tailored for se-

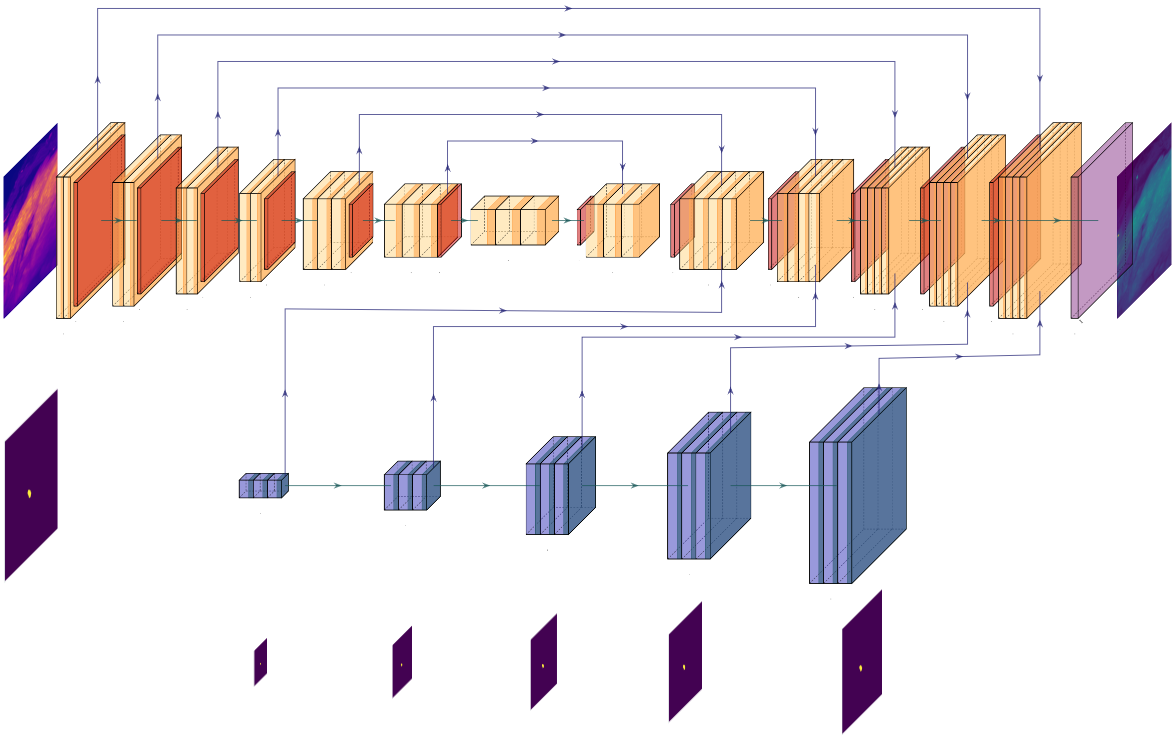

Figure 2. Depiction of the proposed generator architecture, *MetGen*. Our setup processes the input image through a U-net architecture, and the semantic information through Spade ResNet Blocks. The features are concatenated with the final upconvolutional layers and used for spatially and semantically accurate inpainting.

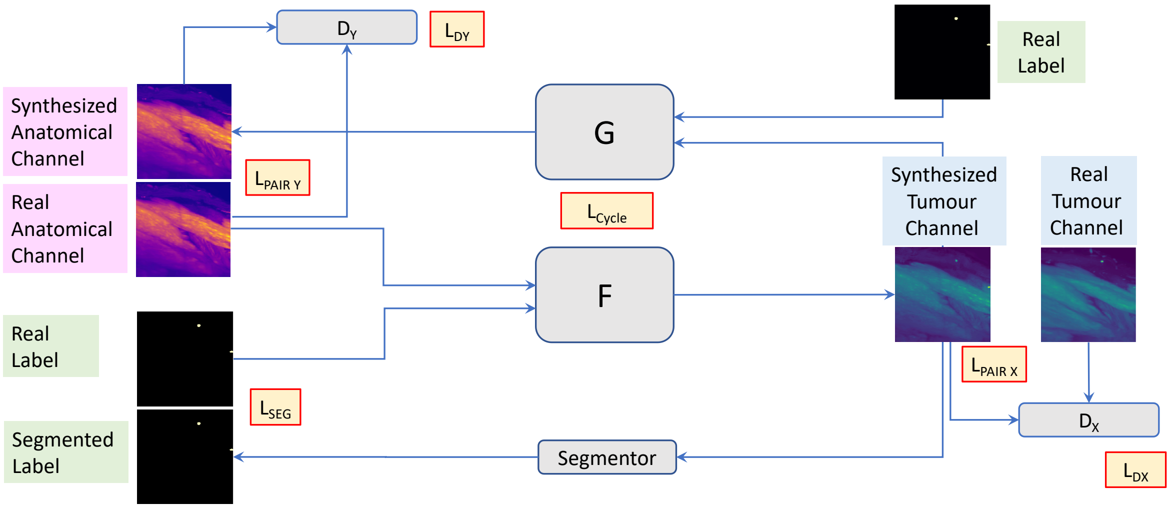

Figure 3. Proposed conditional GAN training setup of *MetGAN*: The generator F learns the mapping from the anatomical to the tumour channel, conditioned on the imposed label, through the discriminator $D_X$. Because we train in a cycle consistent manner, G learns the inverse mapping, through $D_Y$. A pretrained segmentor is used to enforce semantics by punishing F for incorrect tumour placement.

mantic synthesis. Thus, we delimit two paths: one following a traditional U-net architecture [23, 34], that receives the anatomy channel as input; and a second path composed of 7 Spade ResNet Blocks[21], which process the label-input to the network. We then merge the resulting features of the two paths in the upconvolutional layers of the U-net decoder (see Figure 2). We have observed that this separation better preserves the flow of semantic information and results in a more accurate label-based inpainting compared to a naive channel-wise concatenation of the input image and the annotation.

Furthermore, we want to ensure that the generated output is consistent with the desired label - a constraint we enforce through a pre-trained and frozen segmentor network, employed as a passive player in the training process. Its role is twofold: not only does it enforce the placement of metastases at the desired locations, but it also helps to suppress hyperintensities in the anatomy channel which are preserved in the case of CycleGAN, creating semantic ambiguities.
To clarify our terminology, we refer to our generator as seen in Figure 2 as *MetGen*, and to the whole GAN architecture (as seen in Figure 3) as *MetGAN*.

### 2.2. Training Losses

As presented in Figure 3, our final loss consists of four terms. Firstly, we define the discriminator and cycle consistency losses similar to [45]:

$$\begin{aligned} L_D =& L_{D_X} + L_{D_Y} \\ =& \mathbb{E}_x[logD_X(x)] + E_{y,l}[log(1 - D_X(G(y,l)] + \\ & E_y[logD_Y(y)] + E_{x,l}[log(1 - D_Y(F(x,l)] \end{aligned} \tag{1}$$

$$\begin{aligned} L_{Cycle} =& \mathbb{E}_{x \sim p_{(x)}}[\|G(F(x,l),l) - x\|_1] + \\ & \mathbb{E}_{y \sim p_{(y)}}[\|F(G(y,l),l) - y\|_1], \end{aligned} \tag{2}$$

where we denote: X - the anatomical domain, Y - the tumour domain, L - the binary domain marking the presence or absence of a tumour; *x,l* $\sim p_{(x,l)}$ and *y,l*$\sim p_{(y,l)}$ are samples from domain X and Y; $F(x,l)$ is a mapping from X × L → Y; and $G(y,l)$ a mapping from Y × L → X, with $D_X$ and $D_Y$ as corresponding discriminator functions.

Secondly, in order to ensure the desired segmentation map is respected, and that the setup is robust against artefacts and hyperintesities in the autofluorescence channel, we use a weighted binary cross-entropy loss $L_{Segm}$, given by the predictions of the segmentor network, penalizing the discrepancy between the real and segmented label. Lastly, to leverage the paired nature of our data and to facilitate feature adaptation achieved in the domain translation task, we also use a pair-wise loss between real and generated images in each domain:

$$\begin{aligned} L_{Pair} &= L_{Pair_X} + L_{Pair_Y} \\ &= \mathbb{E}_{x,y}[\|F(x,l) - y\|_1] + E_{x,y}[\|G(y,l) - x\|_1]. \end{aligned} \tag{3}$$

The final loss is a weighted linear combination of these components, where the parameters $\alpha_{1-4}$ are hyperparameters adjustable per dataset.

$$L_{final} = \alpha_1 L_D + \alpha_2 L_{Cycle} + \alpha_3 L_{Segm} + \alpha_4 L_{Pair}. \tag{4}$$

## 3. Experiments

We compare our proposed solution to established baseline methods in image translation or semantic synthesis in general computer vision, namely Pix2Pix[11], CycleGAN [45], SPADE [21], and SEAN [46], as well as medical applications: SIFA [5] and RedGAN [22]. Additionally, in an ablation study, we also test our network against mixture models, trained with our generator and/or with an additional segmentation loss. Furthermore, we perform a robustness study exploring our solution's performance under noisy labels. Lastly, we implement a downstream study, where we train segmentation networks on synthetic, real and mixed data.

**Dataset**: In our evaluation, we use a publicly available light sheet microscopy dataset [20, 26]. The dataset contains 1602 300x300 pixel samples, with their ground truth annotations. We put aside 20 % of the data as a test set that is unseen by any network. The test set includes cases with and without metastases, with a balanced organ-based distribution. The samples are resized to the size of 256x256 and normalized to [-1,1]. Random rotations are used at train time. We evaluate the generated images from a qualitative and a quantitative point of view, based on the unseen test set. For this, we use 414 real triplets of anatomical channel, real label, and tumour channel. More information about the architecture components can be seen in Table 2.

**Evaluation Metrics**: We used five image similarity metrics to evaluate the synthetic images: : mean absolute error (MAE), mean sum of squared differences (MSD), image similarity index measure (SSIM), Frechet inception distance (FID) [9], and the learned perceptual image patch similarity (LPIPS) [41]. Among them, FID and LPIPS are two metrics that calculate the distance between the feature vectors obtained from pre-trained deep networks such as [30] for real and synthetic images.

**Implementation details**: We train our models using Pytorch, on an NVIDIA GeForce RTX 2080, for 200 epochs, with batch size=1, using Adam optimizer with $\beta_1$ = 0.5, $\beta_2$ = 0.999, initial learning rate = 0.0002 and linear learning rate decay after 100 epochs. The pre-trained segmentor network (U-net) developed by Pan *et al.* in [20] is loaded, and not modified during training. As a discriminator, we use PatchGAN with 3 layers and a field of view of 70x70 pixels. Training our final method takes approximately 24 hours. For the final loss function, we use $\alpha_1 = 1$, $\alpha_2 = 10$, $\alpha_3 = 100$, and $\alpha_4 = 10$, obtained empirically. The baseline methods were adapted to fit our requirements, and a description of this can be found in the supplementary material.

# 4. Results

## 4.1. Quantitative Results

Table 1. In a quantitative comparison, our proposed method outperforms most state of the art methods in MAE, MSD, SSIM and LPIPS. Our improvements are all significant based on t-test analysis (all p-values $<0.005$). Best scores are indicated in bold digits. Only in one measure (FID) MetGAN is outperformed by one method which is CycleGAN.

| Method | MAE↓ | MSD↓ | SSIM↑ | FID↓ (×10) | LPIPS↓ |
|---|---|---|---|---|---|
| Pix2Pix | 0.122 | 0.028 | 0.650 | 9.364 | 0.239 |
| CycleGAN | 0.129 | 0.031 | 0.656 | **6.731** | 0.223 |
| RedGAN | 0.214 | 0.070 | 0.174 | 48.804 | 0.618 |
| SPADE | 0.346 | 0.190 | 0.372 | 48.262 | 0.569 |
| SPADE+VAE | 0.295 | 0.137 | 0.396 | 44.125 | 0.567 |
| SEAN+VAE | 0.299 | 0.139 | 0.394 | 34.612 | 0.586 |
| SIFA | 0.302 | 0.118 | 0.517 | 26.898 | 0.605 |
| MetGAN (Ours) | **0.111** | **0.023** | **0.700** | 7.945 | **0.214** |

In Table 1, we compare the ”tumour channel” images generated by various methods to the corresponding ground truth using five image similarity metrics. In FID, we achieve close performance when comparing to CycleGAN (7.945 *vs.* 9.364). We observe that *MetGAN* leads to a consistent an statistically significant (p value $< 0.005$) improvement of the generated images compared to baseline according to 4 out of the 5 metrics. Thus, our method obtains a good image similarity, as well as a similar distribution to the ground truth, according to numerical, as well as perceptual metrics.

## 4.2. Qualitative Results

Figure 4 shows examples of qualitative results obtained for our method and state-of-the-art methods. We observe the following common features: while standalone Pix2Pix and CycleGAN obtain good image similarity and domain translation, they either fail to create the desired tumours, or the networks hallucinate features in undesired places. Another Cycle-GAN based approach, SIFA, retains most of the information of the autofluorescence channel and overimposes tumours, but these are not realistic, as they lack depth and dimension. On the other hand, the methods that are specialized for generating structures in the target domain, such as SPADE, SEAN and RedGAN, respect the semantic maps, but they fail at generating diverse and realistic backgrounds. Additionally, when training RedGAN, the network converged to a solution that portrays heavy checkerboard artefacts, an issue we also observed when training ResNet-based generators. On the other hand, *MetGAN* consistently places the tumours at the imposed label location, whilst adapting features from the anatomical domain.

Overall, the combination of our quantitative and qualitative results conclusively proves how MetGAN is generating superior tumour images for our use case.

## 4.3. Additional Experiments

### 4.3.1 Ablation Study

In order to investigate the effects of the components of our setup, we carry out an ablation study on two different datasets: the metastasis dataset described in Section 3, and a second microscopy dataset of the mouse peripheral nervous system. For both cases, the anatomy channel is translated to contrast-enhanced channel, with semantics based on the imposed annotation of metastases or nerves.

**Main Dataset**: We trained combinations of Pix2Pix, CycleGAN and *MetGen* with or without a segmentor and pair-wise loss. Quantitative and qualiative results can be observed in Table 2 and Supplementary Figure 1.We can observe that standalone Pix2Pix and CycleGAN produce good image similarity, but incorrect tumour placement. Adding a segmentor network helps to ensure that the new structures are inpainted at the desired location, but still leads to additional unwanted tumours (Pix2PixSeg), or fails on the domain translation task (CycleGANSeg). The use of our generator improves the qualitative synthesis in both conditional (MetGenCondSeg) and cycle-consistent setups (MetGAN-). Nevertheless, MetGenCondSeg often fails to maintain features from the anatomical channel, producing dark images with bright metastases, which is a simple, but only occasionally correct solution to the task. On the other hand, MetGAN- keeps too many of the original features. An optimal balance is reached with our proposed solution, *MetGAN*, which consistently places tumours at the imposed label location, whilst adapting features from the anatomical domain, and also obtaining the best quantitative metrics.

**Nerve Dataset**: Furthermore, we test our proposed generator (without a segmentor) against the baseline on an anatomically different dataset; the murine peripheral nervous system. This second dataset is similar to the metastases dataset; it has a similar in resolution and contains two-channel (autofluorescence and contrast) images. As the density of structures of interest exceeds that of the metastases by far, we process these samples in a slice-wise manner (unlike the projection-based manner used for the metastases). Therefore, from 18 400x400x400 pixel volumes, we select 10500 images (slices) for training, and 2000 images for testing (25%). For information about the acquisition of the second dataset, please see Cai *et al.* [3].

From both qualitative and quantitative points of view, we can observe that using *MetGen* improves the similarity scores of the conditionally generated images, while respecting the semantics. Unlike for the metastasis dataset, we easily outperform the baseline without the need for an additional segmentor network. We attribute this result to two factors: the larger amount of training data and the reduced

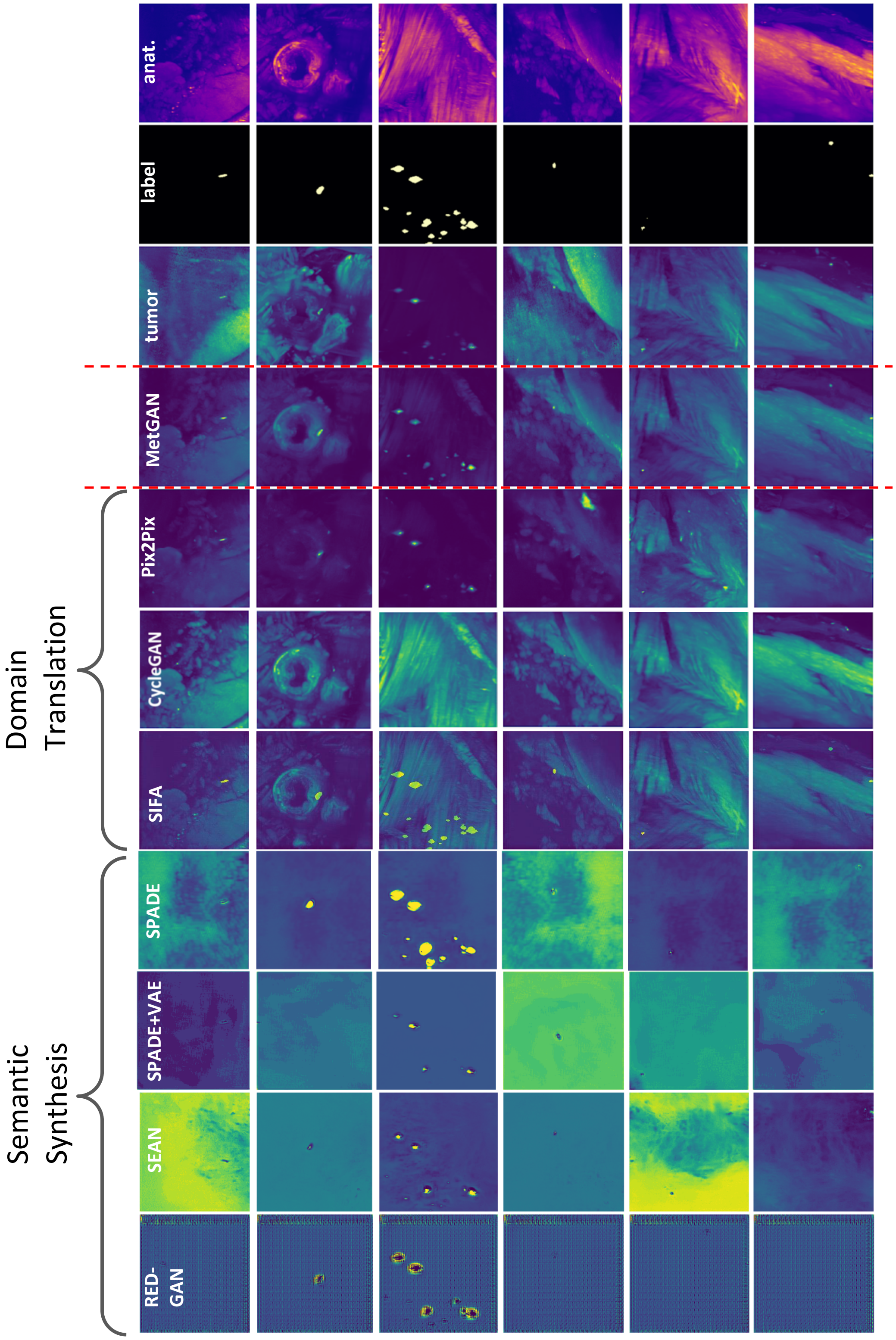

Figure 4. Qualitative results from our proposed generative method compared to the previous state-of-the-art methods. We can see that state of the art domain translation methods fail at respecting the imposed semantic map and/or underperform in the domain translation, whilst semantic synthesis methods fail at generating realistic backgrounds. Our method generates the most realistic looking images, with correct tumour placement.

Table 2. Quantitative comparison on *metastasis dataset*, our proposed method outperforms the baseline and ablated models in MAE, MSD, SSIM and LPIPS. Our improvements are all significant based on t-test analysis (p-values <0.005). Best scores are indicated in bold.

| Name | Generator | Segmentor | Pair Loss | Cycle-Con | MAE↓ | MSD↓ | SSIM↑ | FID↓ (× 10) | LPIPS↓ |
|---|---|---|---|---|---|---|---|---|---|
| Pix2Pix | U-net | | ✓ | | 0.122 | 0.028 | 0.650 | 9.364 | 0.239 |
| Pix2PixSeg | U-net | ✓ | ✓ | | 0.118 | 0.026 | 0.658 | 9.364 | 0.239 |
| CycleGAN | U-net | | | ✓ | 0.129 | 0.031 | 0.656 | **6.731** | 0.223 |
| CycleGANSeg | U-net | ✓ | | ✓ | 0.210 | 0.059 | 0.563 | 26.308 | 0.467 |
| MetGenCond | MetGen | | ✓ | | 0.120 | 0.026 | 0.627 | 10.636 | 0.257 |
| MetGenCondSeg | MetGen | ✓ | ✓ | | 0.125 | 0.028 | 0.643 | 8.938 | 0.227 |
| MetGenCycle | MetGen | | ✓ | ✓ | 0.154 | 0.042 | 0.626 | 10.787 | 0.251 |
| MetGAN - | MetGen | ✓ | | ✓ | 0.120 | 0.026 | 0.654 | 7.450 | 0.216 |
| MetGAN | MetGen | ✓ | ✓ | ✓ | **0.111** | **0.023** | **0.700** | 7.945 | **0.214** |

Table 3. Quantitative validation of our generator, *MetGen*, (without a segmentor) versus the baseline and ablated models, on a second, proprietary dataset pertaining a mouse peripheral nervous system. MAE, MSD, and SSIM are used for comparing generated images to the ground truth. Best scores are indicated in bold digits.

| Name | Generator | Segmentor | Pair Loss | Cycle-Con | MAE↓ | MSD↓ | SSIM ↑ |
|---|---|---|---|---|---|---|---|
| Pix2Pix | U-net | | ✓ | | 0.081 | 0.019 | 0.629 |
| CycleGAN | U-net | | | ✓ | 0.091 | 0.024 | 0.614 |
| MetGen Cond | MetGen | | ✓ | | 0.078 | 0.018 | 0.643 |
| MetGen Cycle | MetGen | | ✓ | ✓ | **0.075** | **0.016** | **0.656** |

inconsistencies between annotations and contrast images.

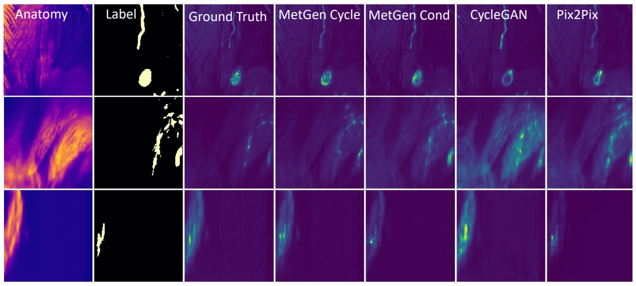

Figure 5. Validation of our generator, *MetGEN* (no segmentor) on a dataset of the mouse peripheral nervous system. The anatomy channel is translated to a contrast-enhanced nerve channel, with semantics based on the imposed annotation. We trained *MetGen* in a Cycle-consistent setup (MetGen-Cycle) or with a conditional discriminator (MetGen Cond). We compare with CycleGAN and Pix2Pix. We can observe that the best similarity between generated and ground truth image is obtained by our setup.

### 4.3.2 Robustness Study

In order to study our model's sensitivity to annotation inconsistencies, we train MetGAN, Pix2Pix and CycleGAN with data consisting of 25% shuffled labels. It can be observed in Figure 6 that our setup is robust to label noise, producing superior results compared to baseline methods, which resort to generating metastases that are randomly placed on the produced images.

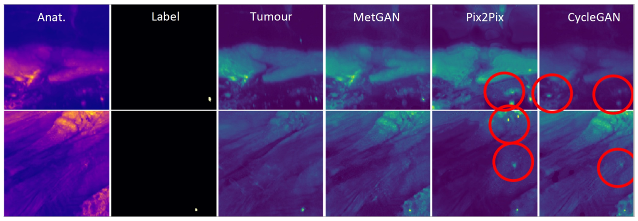

Figure 6. Robustness study. We train *MetGAN*, Pix2Pix and CycleGAN with data containing only 25% shuffled labels. We can observe that MetGAN is robust to label inconsistency, and produces realistic images that are in line with the input.

### 4.3.3 Downstream segmentation task analysis

We analyze how *MetGAN* images can be used to augment real data for training segmentation networks. We train segmentation models on sets of synthetic, real or mixed data, with a varying number of samples from each set. For generating synthetic data, we use randomized unpaired combinations of real anatomical channel images, and real non-zero labels or combinations thereof (e.g. 2 non-zero labels merged together, in order to create more objects of interest in one image). As a segmentor network, we use the U-net developed in [20]. We perform 5-fold cross-validation for all experiments. We evaluate the segmentation performance using lesion-wise Dice, precision, recall and Jaccard Index, on an unseen test dataset (which is not used in GAN training or image generation either). Additionally, segmentation scores of the synthetically generated test set obtained with the original segmentor are provided in the supplementary material.

**Purely synthetic data**: By training a segmentation network only with synthetically generated data, we observe that we achieve a similar performance (78.8% Dice) as by using real data (79.8% Dice). Moreover, for a low number of samples, we not only outperform the model trained with real data, but also offer better training stability with decrease in dataset size. We attribute this to the fact that our approach can generate images with an increased number of objects of interest, allowing us to have a more representative depiction of the data distribution, even in a low sample regime. Nevertheless, increasing the amount of training images past a certain point (≈1000 samples, in our case) results in a slight decrease of performance, as the network overfits on synthetic data.

Table 4. Mean segmentation performance of models trained on real (R), synthetic (S), and mixed data. We see that training based on synthetic data reaches a performance similar to the real data, even outperforming real data at low numbers of training samples. The best segmentation scores can be obtained by combining real and synthetic data.

| Samples | DICE↑ | Prec.↑ | Rec.↑ | J. I.↑ |
|---|---|---|---|---|
| 100 synthetic | 0.657 | 0.607 | 0.718 | 0.489 |
| 120 synthetic | 0.680 | 0.611 | 0.770 | 0.515 |
| 240 synthetic | 0.744 | 0.711 | 0.783 | 0.592 |
| 480 synthetic | 0.770 | 0.767 | 0.776 | 0.626 |
| 1000 synthetic | 0.788 | 0.776 | 0.804 | 0.650 |
| 1800 synthetic | 0.776 | 0.780 | 0.773 | 0.640 |
| 200 real(25%) | 0.470 | 0.465 | 0.538 | 0.307 |
| 280 real(35%) | 0.614 | 0.684 | 0.652 | 0.443 |
| 400 real(50%) | 0.784 | 0.810 | 0.764 | 0.645 |
| 810 real(100%) | 0.790 | 0.800 | 0.781 | 0.653 |
| 200R+500S | 0.809 | 0.809 | 0.811 | 0.679 |
| 400R+500S | 0.819 | 0.812 | **0.828** | 0.693 |
| 810R+500S | **0.826** | **0.842** | 0.811 | **0.704** |
| 810R+1500S | 0.823 | 0.823 | 0.824 | 0.699 |

**Mixed data**: Compared to training the segmentor purely on real images, we observe that our augmentation can increase the performance. Our experiments highlight that, by using as little as 25% of the available real data, together with synthetically generated samples, we can outperform the real data baseline. Adding our synthetic data improves the performance up to a plateau point, where we speculate that the limitations are caused by the inherent variability in annotation quality (see Figure 1). Past this point, adding more generated samples leads to overfitting and a slight decrease in performance. On a per tumour basis, our data augmentation increases the mean number of detected metastases (true positives) from 83 to 95, whilst simultaneously decreasing the number of false positives from 41 to 21. The detection is especially improved for small-sized or dim tumours located in the lungs, showing that our network can produce diverse objects that can be used to improve difficult cases.

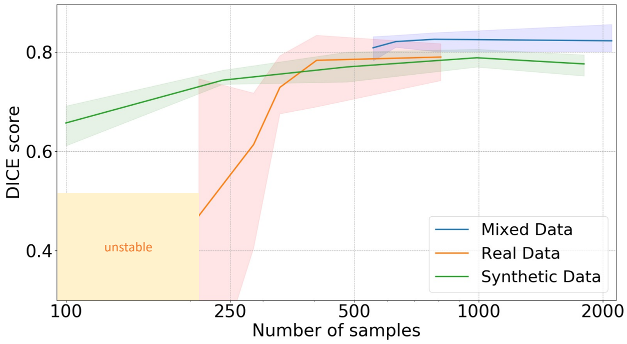

Figure 7. Segmentation performance as function of dataset size during training. Mean values as lines (see Table 4), and minimum and maximum achieved during 5-fold cross validation as delimiters of the areas. We can observe that the inclusion of synthetic data makes training the segmentation network more stable; especially in the case of small datasets, where training on real data is unstable and volatile. On the other hand the performance obtained with purely synthetic or mixed data is more stable.

## 5. Conclusions

In this paper, we introduce a novel generative method, which is able to leverage real anatomical information to generate realistic image-label pairs of tumours. We designed a dual-pathway generator, for the anatomical image and label, trained in a cycle-consistent fashion, which is constrained by an independent, pretrained segmentor. This enables concurrent domain adaptation and semantic synthesis. We generate images which are substantially more realistic in terms of quantitative and qualitative results, compared to different state of the art models; in a manner that is robust to inconsistencies. Moreover, we train segmentation networks on real, generated and mixed data. We find that data synthesized with our method improves segmentation; both from a training stability point of view, observable at low data regimes; as well as from a lesion-detection point of view. Using our method leads to higher segmentation scores when used to augment real data, and can potentially be further exploited by focusing on underrepresented or low-performance cases.

## References

[1] Kumar Abhishek and Ghassan Hamarneh. Mask2lesion: Mask-constrained adversarial skin lesion image synthesis. In *International Workshop on Simulation and Synthesis in Medical Imaging*, pages 71–80. Springer, 2019.

[2] Cher Bass, Tianhong Dai, Benjamin Billot, Kai Arulkumaran, Antonia Creswell, Claudia Clopath, Vincenzo De Paola, and Anil Anthony Bharath. Image synthesis with a convolutional capsule generative adversarial network. In *Proceedings of The 2nd International Conference on Medical Imaging with Deep Learning*, volume 102 of *Proceedings of Machine Learning Research*, pages 39–62. PMLR, 08–10 Jul 2019.

[3] Ruiyao Cai, Chenchen Pan, Alireza Ghasemigharagoz, Mihail Ivilinov Todorov, Benjamin Förstera, Shan Zhao, Harsharan S. Bhatia, Arnaldo Parra-Damas, Leander Mrowka, Delphine Theodorou, Markus Rempfler, Anna L.R. Xavier, Benjamin T. Kress, Corinne Benakis, Hanno Steinke, Sabine Liebscher, Ingo Bechmann, Arthur Liesz, Bjoern Menze, Martin Kerschensteiner, Maiken Nedergaard, and Ali Ertürk. Panoptic imaging of transparent mice reveals whole-body neuronal projections and skull–meninges connections. *Nature Neuroscience*, 2019.

[4] Krishna Chaitanya, Neerav Karani, Christian F Baumgartner, Anton Becker, Olivio Donati, and Ender Konukoglu. Semi-supervised and task-driven data augmentation. In *International conference on information processing in medical imaging*, pages 29–41. Springer, 2019.

[5] Cheng Chen, Qi Dou, Hao Chen, Jing Qin, and Pheng Ann Heng. Unsupervised bidirectional cross-modality adaptation via deeply synergistic image and feature alignment for medical image segmentation. *arXiv preprint arXiv:2002.02255*, 2020.

[6] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018.

[7] Joseph Paul Cohen, Margaux Luck, and Sina Honari. Distribution matching losses can hallucinate features in medical image translation. In *International conference on medical image computing and computer-assisted intervention*, pages 529–536. Springer, 2018.

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[9] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[10] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei A. Efros, and Trevor Darrell. Cycada: Cycle consistent adversarial domain adaptation. In *International Conference on Machine Learning (ICML)*, 2018.

[11] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[12] Qiangguo Jin, Hui Cui, Changming Sun, Zhaopeng Meng, and Ran Su. Free-form tumor synthesis in computed tomography images via richer generative adversarial network. *Knowledge-Based Systems*, 218:106753, 2021.

[13] Salome Kazeminia, Christoph Baur, Arjan Kuijper, Bram van Ginneken, Nassir Navab, Shadi Albarqouni, and Anirban Mukhopadhyay. Gans for medical image analysis. *Artificial Intelligence in Medicine*, page 101938, 2020.

[14] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[15] Hongwei Li, Johannes C Paetzold, Anjany Sekuboyina, Florian Kofler, Jianguo Zhang, Jan S Kirschke, Benedikt Wiestler, and Bjoern Menze. Diamondgan: unified multimodal generative adversarial networks for mri sequences synthesis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 795–803. Springer, 2019.

[16] Dongnan Liu, Donghao Zhang, Yang Song, Fan Zhang, Lauren O'Donnell, Heng Huang, Mei Chen, and Weidong Cai. Pdam: A panoptic-level feature alignment framework for unsupervised domain adaptive instance segmentation in microscopy images. *IEEE Transactions on Medical Imaging*, 2020.

[17] Yingzi Liu, Yang Lei, Tonghe Wang, Yabo Fu, Xiangyang Tang, Walter J Curran, Tian Liu, Pretesh Patel, and Xiaofeng Yang. Cbct-based synthetic ct generation using deep-attention cyclegan for pancreatic adaptive radiotherapy. *Medical physics*, 47(6):2472–2483, 2020.

[18] Giovanni Mariani, Florian Scheidegger, Roxana Istrate, Costas Bekas, and Cristiano Malossi. Bagan: Data augmentation with balancing gan. *arXiv preprint arXiv:1803.09655*, 2018.

[19] Johannes C Paetzold, Oliver Schoppe, Rami Al-Maskari, Giles Tetteh, Velizar Efremov, Mihail I Todorov, Ruiyao Cai, et al. Transfer learning from synthetic data reduces need for labels to segment brain vasculature and neural pathways in 3d. In *International Conference on Medical Imaging with Deep Learning–Extended Abstract Track*, 2019.

[20] Chenchen Pan, Oliver Schoppe, Arnaldo Parra-Damas, Ruiyao Cai, Mihail Ivilinov Todorov, Gabor Gondi, Bettina von Neubeck, Nuray Böğürcü-Seidel, Sascha Seidel, Katia Sleiman, et al. Deep learning reveals cancer metastasis and therapeutic antibody targeting in the entire body. *Cell*, 179(7):1661–1676, 2019.

[21] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2337–2346, 2019.

[22] Ahmad B Qasim, Ivan Ezhov, Suprosanna Shit, Oliver Schoppe, Johannes C Paetzold, Anjany Sekuboyina, Florian Kofler, Jana Lipkova, Hongwei Li, and Bjoern Menze. Red-gan: Attacking class imbalance via conditioned generation. yet another medical imaging perspective. *Proceedings of Machine Learning Research*, 1:13, 2020.

[23] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

[24] Veit Sandfort, Ke Yan, Perry J Pickhardt, and Ronald M Summers. Data augmentation using generative adversarial networks (cyclegan) to improve generalizability in ct segmentation tasks. *Scientific reports*, 9(1):1–9, 2019.

[25] Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *International conference on information processing in medical imaging*, pages 146–157. Springer, 2017.

[26] Oliver Schoppe, Chenchen Pan, Javier Coronel, Hongcheng Mai, Zhouyi Rong, Mihail Ivilinov Todorov, Annemarie Müskes, Fernando Navarro, Hongwei Li, Ali Ertürk, et al. Deep learning-enabled multi-organ segmentation in whole-body mouse scans. *Nature communications*, 11(1):1–14, 2020.

[27] Suprosanna Shit, Johannes C Paetzold, Anjany Sekuboyina, Ivan Ezhov, Alexander Unger, Andrey Zhylka, Josien PW Pluim, Ulrich Bauer, and Bjoern H Menze. cldice-a novel topology-preserving loss function for tubular structure segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16560–16569, 2021.

[28] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):60, 2019.

[29] Etsuo A Susaki, Kazuki Tainaka, Dimitri Perrin, Fumiaki Kishino, Takehiro Tawara, Tomonobu M Watanabe, Chihiro Yokoyama, Hirotaka Onoe, Megumi Eguchi, Shun Yamaguchi, et al. Whole-brain imaging with single-cell resolution using chemical cocktails and computational analysis. *Cell*, 157(3):726–739, 2014.

[30] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[31] Mihail Ivilinov Todorov, Johannes Christian Paetzold, Oliver Schoppe, Giles Tetteh, Suprosanna Shit, Velizar Efremov, Katalin Todorov-Völgyi, Marco Düring, Martin Dichgans, Marie Piraud, et al. Machine learning analysis of whole mouse brain vasculature. *Nature Methods*, 17(4):442–449, 2020.

[32] A. Tomczak, S. Ilic, G. Marquardt, T. Engel, F. Forster, N. Navab, and S. Albarqouni. Multi-task multi-domain learning for digital staining and classification of leukocytes. *IEEE Transactions on Medical Imaging*, pages 1–1, 2020.

[33] Hiroki R Ueda, Ali Ertürk, Kwanghun Chung, Viviana Gradinaru, Alain Chédotal, Pavel Tomancak, and Philipp J Keller. Tissue clearing and its applications in neuroscience. *Nature Reviews Neuroscience*, pages 1–19, 2020.

[34] Joris van Vugt. pytorch-unet. *GitHub*, 2019. `https://github.com/jvanvugt/pytorch-unet`.

[35] Simon Vandenhende, Bert De Brabandere, Davy Neven, and Luc Van Gool. A three-player gan: generating hard samples to improve classification networks. In *16th International Conference on Machine Vision Applications (MVA)*, pages 1–6. IEEE, 2019.

[36] Eric Wu, Kevin Wu, D. Cox, and William Lotter. Conditional infilling gans for data augmentation in mammogram classification. *ArXiv*, abs/1807.08093, 2018.

[37] Bingyu Xin, Yifan Hu, Yefeng Zheng, and Hongen Liao. Multi-modality generative adversarial networks with tumor consistency loss for brain mr image synthesis. In *The IEEE International Symposium on Biomedical Imaging (ISBI)*, 2020.

[38] Zhenghua Xu, Chang Qi, and Guizhi Xu. Semi-supervised attention-guided cyclegan for data augmentation on medical images. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 563–568. IEEE, 2019.

[39] Heran Yang, Jian Sun, Aaron Carass, Can Zhao, Junghoon Lee, Zongben Xu, and Jerry Prince. Unpaired Brain MR-to-CT Synthesis Using a Structure-Constrained CycleGAN. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 174–182, Cham, 2018. Springer International Publishing.

[40] Xin Yi, Ekta Walia, and Paul Babyn. Generative adversarial network in medical imaging: A review. *Medical image analysis*, 58:101552, 2019.

[41] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[42] Zizhao Zhang, Lin Yang, and Yefeng Zheng. Translating and segmenting multimodal medical volumes with cycle-and shape-consistency generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9242–9251, 2018.

[43] Sicheng Zhao, Bo Li, Xiangyu Yue, Yang Gu, Pengfei Xu, Runbo Tan, Hu, Hua Chai, and Kurt Keutzer. Multi-source domain adaptation for semantic segmentation. In *Advances in Neural Information Processing Systems*, 2019.

[44] Shan Zhao, Mihail Ivilinov Todorov, Ruiyao Cai, Rami AI-Maskari, Hanno Steinke, Elisabeth Kemter, Hongcheng Mai, Zhouyi Rong, Martin Warmer, Karen Stanic, et al. Cellular and molecular probing of intact human organs. *Cell*, 180(4):796–812, 2020.

[45] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

[46] Peihao Zhu, Rameen Abdal, Yipeng Qin, and Peter Wonka. Sean: Image synthesis with semantic region-adaptive nor-

malization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.