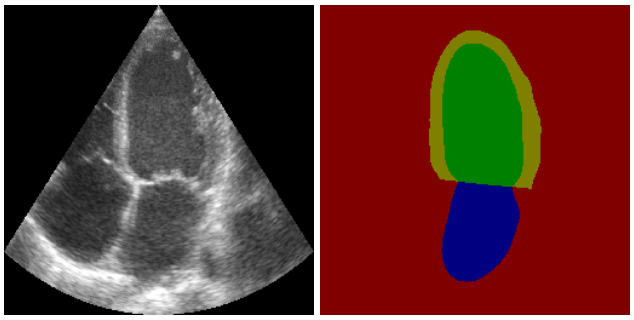
CARDIGAN: TOWARDS BETTER ECHOCARDIOGRAM SEGMENTATION BY DATA GENERATION USING SPADE

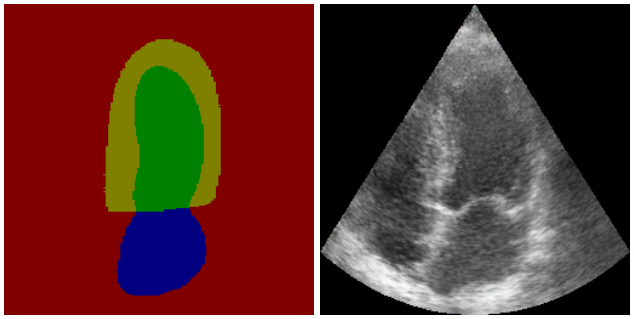
Ulysse Rançon[†] *Corentin Vannier*[†] *Olivier Bernard*[†⋆]

[†] INSA Lyon, Department of Electrical Engineering
[⋆] University of Lyon, CREATIS, CNRS UMR5220, Inserm U1044



(a) A 4-channel end-systole echocardiogram from CAMUS

(b) Its corresponding manually labeled mask



(c) An artificial mask by Nathan Painchaud

(d) Its corresponding SPADE generated echocardiogram

Fig. 1. Using the CAMUS dataset (1a, 1b), we trained a Generative Adversarial Network to create realistic echocardiograms from semantic masks. Feeding it with artificial masks (1c) allowed us to generate a large number of images (1d) and augment the CAMUS dataset.

ABSTRACT

Over the last decade, deep learning allowed major breakthroughs in the field of image processing. Complex models have tackled image classification and segmentation [1, 2] tasks extremely efficiently, even outperforming human experts in some cases.

However, those models often require a lot of training data, which can be hard to obtain in medical imaging, thus limiting their performances in tasks such as echocardiogram segmentation. While the rising capabilities of deep generative models such as Generative Adversarial Networks (GANs) hint at data generation as a workaround to these limitations, this approach is rarely seen in the field of medical imaging.

We decided to train a GAN model using NVIDIA's SPADE[3] on a small echocardiogram dataset[4] and use it augment that dataset. We then trained semantic segmentation models on the augmented dataset, and found that performances were lower than on models trained only on the original dataset. We investigated the reasons of this failure and propose several hypotheses as to why that might be.

The Github repository associated with this paper can be found at `github.com/urancon/cardiGAN`.

Index Terms— Echocardiography, Generative adversarial networks, Image segmentation

1. INTRODUCTION

Echocardiography is a medical imaging technique allowing the detection of heart diseases by the observation of cardiac tissues. Blood ejection fraction is a common metric in the diagnosis of such conditions, and is indirectly estimated by measuring the left ventricular volume at two different times, known as end-diastole (ED) and end-systole (ES).

Semantic segmentation addresses such measurement of cavity volumes, by labeling each pixel of an input echocardiogram to a class (in this case, either myocardium, left ventricle or left atrium). Several years after the proposition of the U-Net [1] and ENet [2] architectures, both of which excel at this particular task, two-dimensional semantic segmentation is on the verge of being considered as a "solved problem".

Data augmentation is known to help machine learning models generalize and perform better at certain tasks. First investigated through conventional image processing (rotation, scaling, filtering, noising...), it has become more and more sophisticated since the introduction of deep generative models such as Variational Auto-Encoders (VAEs) and Generative Adversarial Networks (GANs). Both aim at capturing a statistical distribution of a dataset, and recreating images from

this distribution, seen as points in high-dimensional space.

Conditional image synthesis is a generative task where the model is given some information regarding how it is supposed to construct the output image. Such information can take many forms. In the case of image-to-image translation, the given information is an input image. This is the case of NVIDIA's pix2pixHD model [5]. Another form of conditional image synthesis is semantic image synthesis, when the given information is a semantic map defining the topology of the generated image. In this domain, NVIDIA researchers made a major breakthrough by proposing a GAN model using a new normalization technique called SPADE, capable of excellent performances [3].

2. MATERIALS AND METHODS

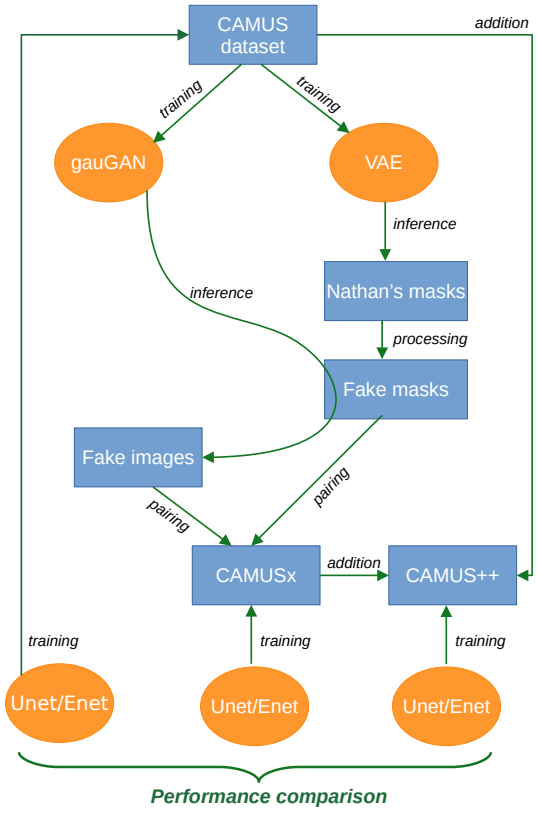


Fig. 2. Our work pipeline. Blue squares represent datasets, while orange circles represent deep learning models.

Figure 2 presents our work in its globality to provide a better understanding of the overall project. Please note that VAE training and inference was done in [7].

2.1. Datasets

CAMUS is an echocardiography segmentation dataset consisting of clinical exams from 500 patients [4]. Data from each patient is composed of 4-chambers and 2-chambers images acquired at both the end-diastole and end-systole instants of the cardiac cycle. As a result, CAMUS dataset comprises a total of 2000 images manually annotated by an expert. Ground-truth annotation images will be further referred as *masks* or *semantic maps*. Data is split between training, validation and test sets, each respectively composed of 1600, 200 and 200 annotated images.

We also had access to 5000 standalone semantic masks generated by Nathan Painchaud using a VAE trained on CAMUS [7]. We started by removing degenerated masks, and ended up with 4634 masks. Then, since their shape and position slightly differed from those of CAMUS masks, we preprocessed them so that the distributions matched. Please refer to annex F for an illustration of how we adapted those distributions. These artificial masks served as conditional inputs to our GAN model for generating realistic echocardiograms.

2.2. Echocardiogram generation

We used the official SPADE implementation [3] that is freely available on GitHub at `github.com/NVlabs/SPADE`. All of our SPADE models have the same pix2pix backbone with 48 convolutional filters for both the generator and the discriminator, and an input size of 256×256.

One barrier inherent to GANs is the difficulty of quantifying their performances, as no objective and comprehensive metric has been found for this purpose. In the case of artificial echocardiogram generation, it is hard to define an objective function that takes into account criteria such as the anatomical plausibility and the presence of typical echographic textures. Thus, the quality of the images generated by our model was manually appreciated by experts in ultrasound imaging. Our evaluation protocol consisted in comparing the generated echocardiograms with their input masks, as well as the corresponding echocardiogram reference from the CAMUS dataset. Compliance with the morphology of the semantic map, the anatomy of the heart and the physics of echocardiography were all important factors for model evaluation.

2.3. Echocardiogram segmentation

We trained three pairs of U-Net and ENet segmentation models with identical hyperparameters. The first pair, referred to as **CAMUS**, was trained on the original CAMUS training set alone to serve as a reference for performance evaluation. The second one, **CAMUS++**, was trained on an augmented version of the CAMUS training set, comprising its 1600 original images plus the 4634 artificial images with their corresponding masks. Finally, we trained a third pair of models only on

		Dice		HD (mm)		ASSD (mm)	
		Endo	**Epi**	**Endo**	**Epi**	**Endo**	**Epi**
CAMUS (2000 images)	**U-Net**	**0.937**	**0.957**	5.0	5.4	1.4	1.5
	ENet	0.939	**0.957**	4.7	5.0	1.3	1.5
CAMUS++ (2000 + 4634 images)	**U-Net**	0.919	0.942	6.1	7.2	1.7	2.0
	ENet	0.932	0.95	4.8	5.7	1.4	1.8
CAMUSx (4634 images)	**U-Net**	0.919	0.942	6.1	7.2	1.7	2.0
	ENet	0.732	0.8	15.2	18.6	6.2	7.9
Wei et al. [6]		0.929	0.955	**4.6**	**4.9**	**1.4**	**1.6**

Table 1. Evaluation results of semantic segmentation models.

the artificial images, **CAMUSx**. For validation and testing, we used the original sets from CAMUS for all three pairs.

Evaluation metrics used on those models are the Dice Coefficient, the Hausdorff Distance (HD) and the Average Symmetric Surface Distance (ASSD). Dice is a popular metric in segmentation tasks, and is directly linked to the most commonly known Intersection over Union (IoU). HD and ASSD are more specific to the medical imaging community. HD is sensitive to outliers and is used to find the largest distance between ground and predicted masks, while ASSD represents the average distance between both sets.

3. EXPERIMENTS

3.1. Training GAN models

To try and obtain the best results, we first trained models with the recommended architecture on the SPADE repository (50 epochs, a batch size of 1, and an ADAM optimizer with $\beta_1 = 0$, $\beta_2 = 0.999$ and a learning rate of 0.0002 that linearly decays starting from the 30^{th} epoch). Then, we tuned some hyperparameters:

The parameter with the most prominent role was the the training set. We started experimenting with small subsets of CAMUS before training on the whole set. The smaller the subsets we used, the more we could see small checkerboard artifacts in the upper corner of the generated images. These artifacts are known to be the result of deconvolution layers [8]. The reason as to why they are especially present in this area remains unanswered, but it may be related to the physics of echography acquisition. We included inferences of models trained on different subsets of CAMUS in appendix C.

The batch size had a similar impact on the quality of the generated echocardiograms. Models trained with smaller batch sizes tended to have more visible checkerboard artifacts. However, we found the images generated by models trained with large batch sizes to be somewhat more flat, with less contrast and less realistic artifacts of heart structures such as papillary muscles. We suspect that this is due to an averaging effect of batch normalization in GANs. We found using a smaller batch size on the largest possible set of data to be a good compromise between checkerboard artifacts and lack of contrast. For illustrations of models trained with different batch sizes, please refer to appendix D.

Finally, one important setting of our GAN model was the presence of a VAE head in the backbone for multimodal synthesis, i.e. within-model variability. Simply put, a VAE head allows the GAN model to generate multiple artificial images for a single semantic map using style images. We tried implementing it, however results were not convincing. Indeed, output images for one mask only had slight brightness and contrast differences. We presume that the model finds a Nash equilibrium in which the generator produces mostly identical outputs and has little interest in deviating from them. This may be caused by simplicity of the CAMUS dataset (small amount of data, and little variability between images). We have included illustrations of a model trained with a VAE head in appendix E. You can see that despite the differences in brightness, contrast, and artifacts in the input style images, there is little to no difference in the generated echocardiograms.

From these observations, we decided to train our models on all 1800 images of both the CAMUS training and validation sets, with a batch size of 4, and without a VAE head.

3.2. Augmenting the original dataset

We ran inference on the aforementioned model using our 4634 artificial masks to generate realistic echocardiograms. We added the resulting 4634 pairs of artificial masks and artificial echocardiograms to CAMUS to create the augmented dataset we referred to as **CAMUS++**.

3.3. Training segmentation models

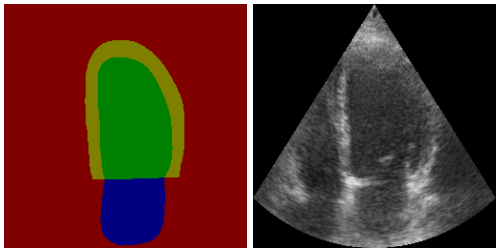
We trained the three pairs of U-Net and ENet models mentionned in subsection 2.3 with identical hyperparameters (40 epochs, a batch size of 8 and a ADAM optimizer using a learning rate of 0.001).

4. RESULTS

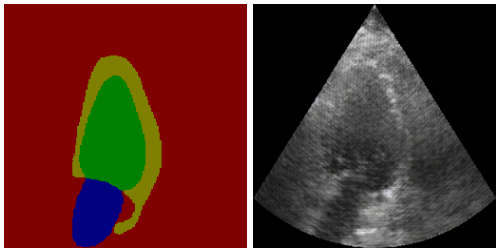
4.1. Echocardiogram generation

Even to the eye of an expert, the artificial data generated by our model are hardly distinguishable from real echocardiograms. Even though it has its own signature such as minor checkerboard patterns or blur, it visually succeeds at capturing the style of the CAMUS dataset, and complies both with the semantic input and the human anatomy. Our model also succeeded in learning common echocardiography artifacts resulting from the physics of ultrasound imaging, as well as sparsely represented heart structures such as papillary muscles (cf. figures 3a, 3b).

In addition, our model is also robust against wrong conditioning: when presented with an anatomically incorrect semantic map, it compensates by covering the implausible regions with believable noise and artifacts (cf. figure 3c, 3d).



(a) Plausible input mask (b) Output image



(c) Degenerate mask (d) Output image

Fig. 3. Example output images from our GAN model. (3a, 3b) illustrate the generation of plausible heart structures. (3c, 3d) illustrate the covering implausible regions with noise.

4.2. Echocardiogram segmentation

The results of our training can be found in table 1. Despite the seemingly positive results of our GAN model, our U-Net and ENet models trained on the augmented dataset show no performance improvement over the ones trained only on CAMUS. Worse yet, performance is actually poorer. This means we did not reach our initial goal of surpassing current models [6] by training well-known segmentation architectures on GAN-augmented data.

5. DISCUSSION

We propose several hypotheses as to why we did not see any improvement in performances, the most obvious of which being that artificial data, either the VAE-generated masks or the GAN-generated images, is biased. In other terms, it could be distant from the original data in the high-dimension original space. This distance would be explained by consistent differences in the generated data, even if imperceptible to the human eye.

We examine those hypotheses through UMAP, a nonlinear dimensionality-reduction algorithm, by projecting real and artificial data from their space of origin to a 3D space. Known for its excellence, UMAP comes from a pure mathematics and topological setting, and claims to have better performances than the formerly used t-SNE for high-dimensional data visualization [9]. As suspected, we can see in the figures of appendix B that the distribution of the artificial data significantly underrepresents some areas the original distribution, while also falling outside of the original distribution in some areas.

Also, from the point of view of information theory, we wonder if it is actually possible to achieve better performances on a dataset by augmenting it with artificial data produced from this same dataset. Although information generation should not be possible, some studies have succeeded in applying strategies similar to ours [10, 11, 12].

As a conclusion, we have not been able to improve segmentation performances by augmenting a dataset by GAN-enhanced conditional echocardiogram generation. As such, our work is a direct answer to [13], a similar work also conducted on CAMUS dataset. Such failure could hint at the impossibility to generate relevant additional data *via* Generative Adversarial Networks However, it should not be taken as such, as the scientific community seems to be able to efficiently train machine learning models on AI-generated data. Our results remain extremely interesting in that we do not have a definitive answer to our problem. In this sense, it appears necessary to further explore the questions we have raised. A first step in this direction would be to compare our approach with successful ones, with the eyes of a confirmed information theorist.

6. ACKNOWLEDGMENTS

We would like to deeply thank Olivier Bernard, the instigator of this research project, for his precious advice and his supervision. Our thanks also go to Nathan Painchaud for his mask generation program.

Appendices

A. REFERENCES

[1] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," *arXiv e-prints*, May 2015.

[2] Adam Paszke, Abhishek Chaurasia, Sangpil Kim, and Eugenio Culurciello, "ENet: A Deep Neural Network Architecture for Real-Time Semantic Segmentation," *arXiv e-prints*, June 2016.

[3] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu, "Semantic Image Synthesis with Spatially-Adaptive Normalization," *arXiv e-prints*, Mar. 2019.

[4] Sarah Leclerc, Erik Smistad, João Pedrosa, Andreas Østvik, Frederic Cervenansky, Florian Espinosa, Torvald Espeland, Erik Andreas Rye Berg, Pierre-Marc Jodoin, Thomas Grenier, Carole Lartizien, Jan D'hooge, Lasse Lovstakken, and Olivier Bernard, "Deep Learning for Segmentation using an Open Large-Scale Dataset in 2D Echocardiography," *arXiv e-prints*, Aug. 2019.

[5] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro, "High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs," *arXiv e-prints*, Nov. 2017.

[6] Hongrong Wei, Hen Cao, Yiqin Cao, Yongjin Zhou, Wufeng Xue, Dong Ni, and Shuo Li, "Temporal-consistent segmentation of echocardiography with co-learning from appearance and shape," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, Cham, 2020, pp. 623–632, Springer.

[7] Nathan Painchaud, Youssef Skandarani, Thierry Judge, Olivier Bernard, Alain Lalande, and Pierre-Marc Jodoin, "Cardiac MRI Segmentation with Strong Anatomical Guarantees," *arXiv e-prints*, July 2019.

[8] Augustus Odena, Vincent Dumoulin, and Chris Olah, "Deconvolution and checkerboard artifacts," *Distill*, Oct. 2016.

[9] Leland McInnes, John Healy, and James Melville, "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction," *arXiv e-prints*, Feb. 2018.

[10] Maayan Frid-Adar, Idit Diamant, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan, "GAN-based Synthetic Medical Image Augmentation for increased CNN Performance in Liver Lesion Classification," *arXiv e-prints*, Mar. 2018.

[11] Yu-Xiong Wang, Ross Girshick, Martial Hebert, and Bharath Hariharan, "Low-Shot Learning from Imaginary Data," *arXiv e-prints*, Jan. 2018.

[12] Antreas Antoniou, Amos Storkey, and Harrison Edwards, "Data Augmentation Generative Adversarial Networks," *arXiv e-prints*, Nov. 2017.

[13] Amir H. Abdi, Teresa Tsang, and Purang Abolmaesumi, "GAN-enhanced Conditional Echocardiogram Generation," *arXiv e-prints*, Nov. 2019.

[14] Kathuria Ayoosh, "Understanding GauGAN," *Paperspace*, Jan. 2020.

[15] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo, "MaskGAN: Towards Diverse and Interactive Facial Image Manipulation," *arXiv e-prints*, July 2019.

B. UMAP PROJECTION OF CAMUS AND CAMUSX

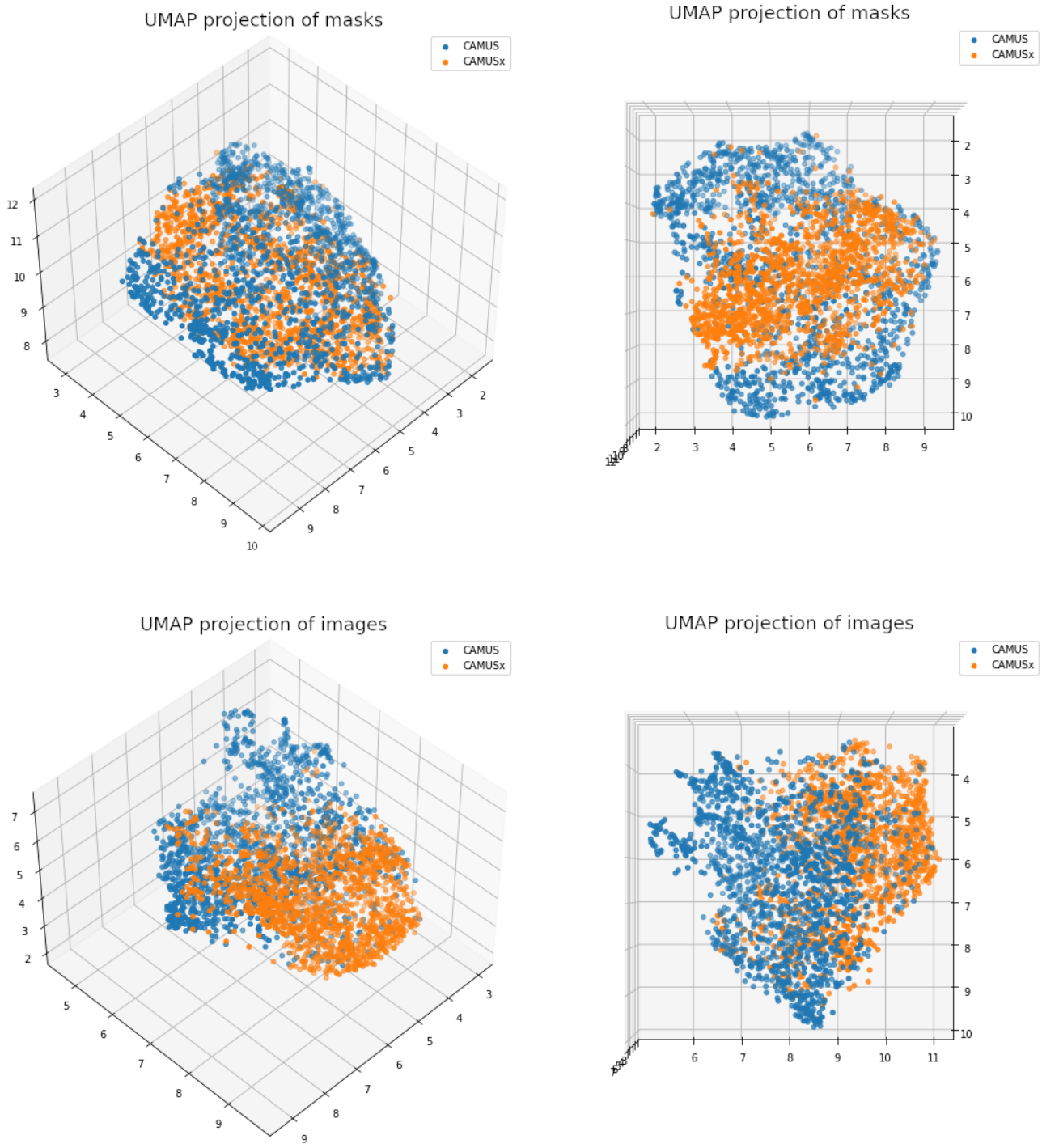
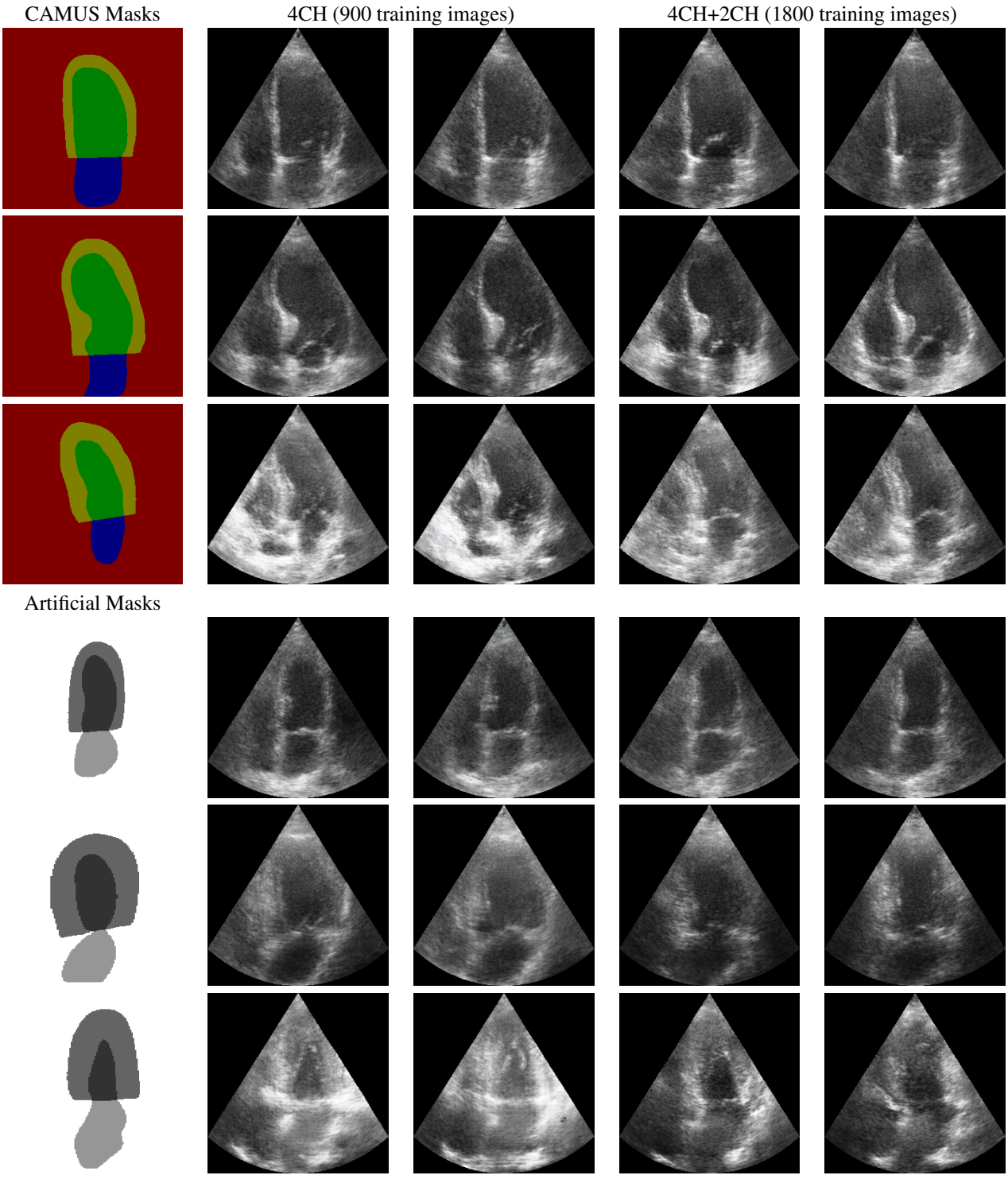


Fig. 4. UMAP projections of CAMUS and CAMUSx masks and images

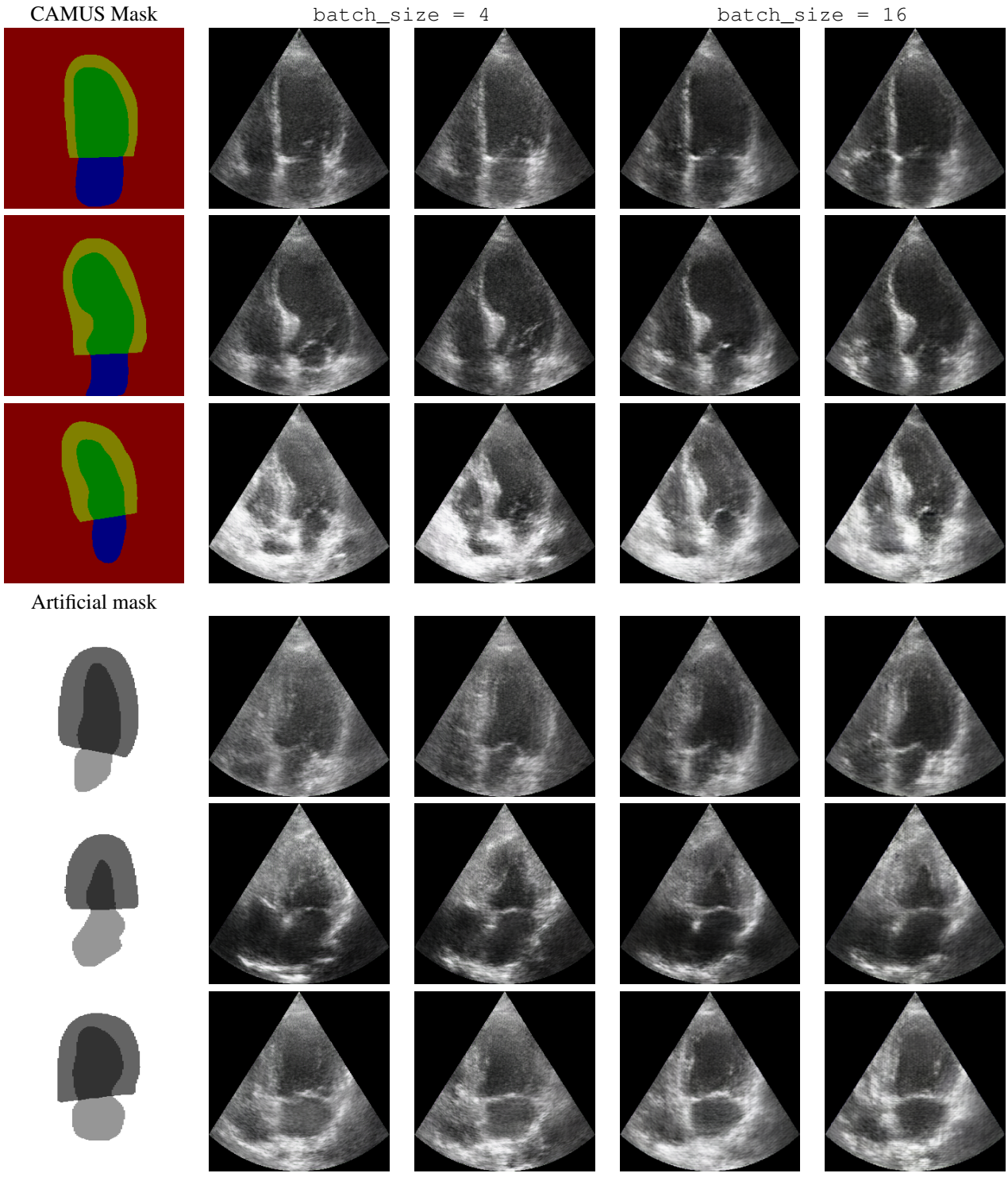
C. INFLUENCE OF DATASET SIZE

Table 2. Inferences by two networks trained on CAMUS 4-chambers images and two networks train on both 4-chambers and 2-chambers images, all of them using a batch size of 4.



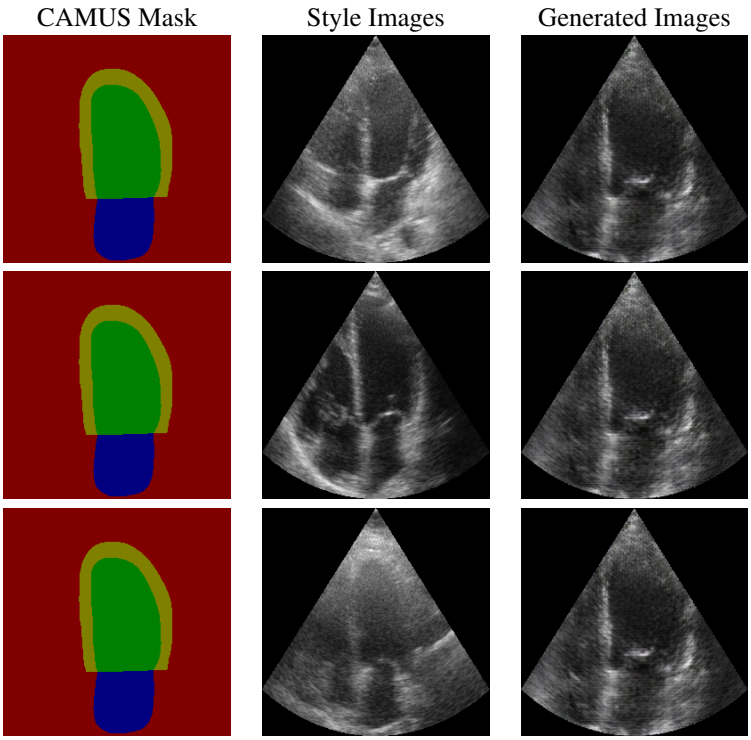
D. INFLUENCE OF BATCH SIZE

Table 3. Inferences by two networks trained using a batch size of 4 and two networks trained using a batch size of 16, all of them on CAMUS 4-chambers images (900 training images).



E. INFLUENCE OF VAE HEAD

Table 4. Inferences by one network trained with a VAE head on a single mask with different style images.



F. PREPROCESSING OF ARTIFICIAL MASKS

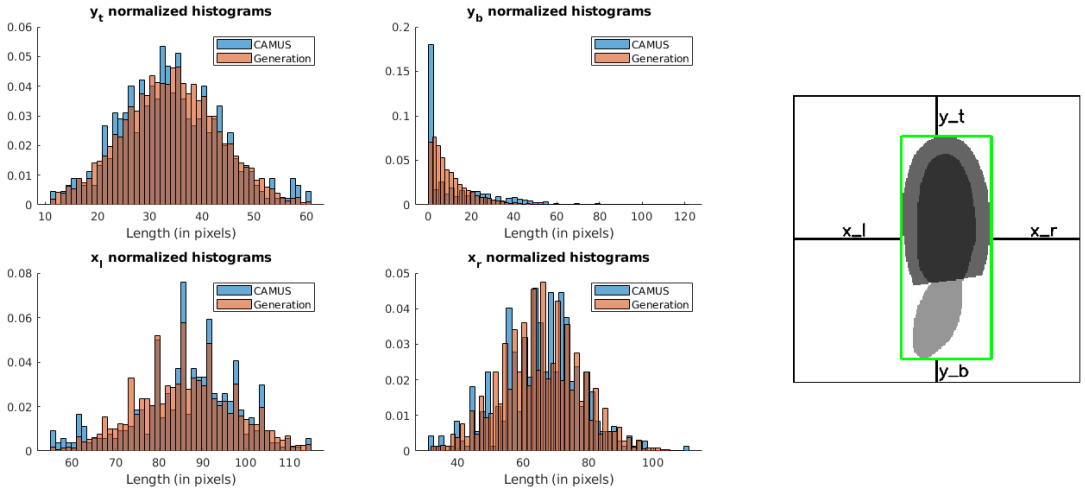


Fig. 5. Preprocessing of Nathan Painchaud's masks