

Supplementary Material

Learning shape distributions from large databases of healthy organs: applications to zero-shot and few-shot abnormal pancreas detection

Rebeca Vétel, Clément Abi-Nader, Alexandre Bône, Marie-Pierre Vullierme, Marc-Michel Rohé, Pietro Gori, and Isabelle Bloch

A. Model Architecture

As detailed in Figure 1, the proposed variational autoencoder (VAE) followed a 3D encoder-decoder architecture. The network topology (number of convolutions per block, filter sizes) was chosen based on the nnU-Net self-configuring procedure [4]. The model was trained on 800 epochs. We used the Stochastic Gradient Descent optimizer with an initial learning rate of 10^{-4} following a polynomial decay. To avoid memory issues and mimic larger batch sizes, better suited for VAEs [5], we used a batch size of 8 with a gradient accumulation step of 5.

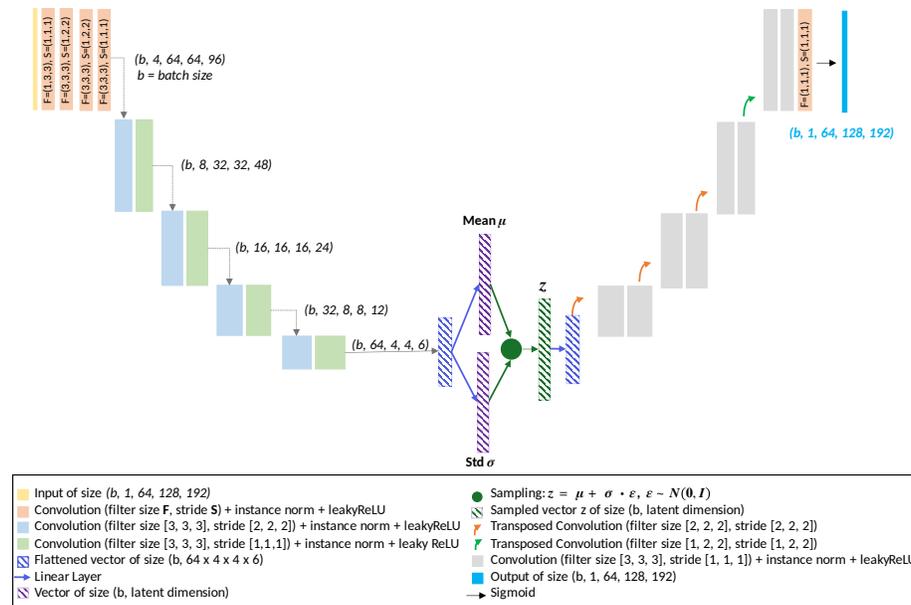


Fig. 1. Architecture of the proposed VAE

B. Inclusion criteria

The training data were selected from a dataset consisting of 2606 portal phase abdominal CT scans acquired from patients with potential liver disease. Each scan was accompanied by a radiological report as well as segmentations of the liver and liver lesions, done manually by radiologists. In order to exclude any patient whose condition could affect the pancreas, different exclusion criteria were applied. First, all patients with metastatic cancer or lesions reported as "unevaluable" were excluded. Secondly, in order to avoid overly large liver lesions, patients with lesion volumes exceeding the 99th percentile were also excluded. Thirdly, cases in which lesions protruded from the liver were also removed from the study. Finally, the pancreas segmentation masks were automatically generated for the remaining patients, and only the cases for which the mask consisted of one single connected component were retained, this last criterion acting as a quality control on the segmentation masks. In the end, 1200 cases were retained to build the healthy pancreas database for training.

Table 1. Individual splits of the different training sets \mathcal{D}_N , as well as their corresponding inclusion criteria.

Name	N_{tot}	N_{train}	N_{val}	Liver lesion			Other abdominal pathology
				Benign	Non suspicious	Suspicious	
\mathcal{D}_{300}	300	240	60	✓	✗	✗	✗
\mathcal{D}_{600}	600	480	120	✓	✓	✗	✗
\mathcal{D}_{900}	900	720	180	✓	✓	✓	✗
\mathcal{D}_{1200}	1200	960	240	✓	✓	✓	✓

C. Reconstruction results

As an additional indicator for the good reconstruction capacity of the proposed model, we also reported, in Table 2, the Dice scores computed between the input and output of the VAE for the 244 subjects of \mathcal{D}^{Test} . As for the AUC scores obtained in zero-shot AD, the Dice scores increased with larger latent space dimensions and larger training sets. In particular, best reconstruction performances coincided with the best detection performances, obtained on the model trained on \mathcal{D}_{1200} with a latent dimension $L = 1024$.

D. Comparison with other SSM methods

We compared our framework with two state-of-the-art methods: active shape models (ASM) [2] and Large Deformation Diffeomorphic Metric Mapping (LDDMM) using the Deformetrica software [1]. Concerning ASM, we computed the signed distance map of the pancreas 3D contours of each subject. For LDDMM,

Table 2. Results for segmentation masks reconstruction. For each experiment, corresponding to a specific training size and latent space dimension, we report the mean and standard deviation of Dice scores obtained by bootstrapping with 10000 repetitions. Best results by line are underlined and by column are in **bold**.

	L_{16}	L_{64}	L_{256}	L_{1024}
\mathcal{D}_{300}	73.7 \pm 7.2	79.9 \pm 5.7	80.9 \pm 7.4	<u>81.5 \pm 7.4</u>
\mathcal{D}_{600}	75.5 \pm 7.5	83.4 \pm 5.4	85.0 \pm 5.8	83.9 \pm 5.9
\mathcal{D}_{900}	78.1 \pm 5.9	85.0 \pm 5.8	<u>86.1 \pm 5.4</u>	85.4 \pm 5.8
\mathcal{D}_{1200}	77.3 \pm 6.4	85.9 \pm 5.4	86.2 \pm 5.3	88.5 \pm 4.4

we estimated a Bayesian Atlas [3] parametrized by 576 control points. In both cases, we performed a PCA on the shape-encoding parameters to obtain a latent vector of dimension 1024 for each subject. We compared them with our method also using a latent dimension $L = 1024$. All methods were trained on \mathcal{D}_{1200} . Results, reported in Table 3, show that our method outperforms ASM and LDMM when the number of training samples is small (Train/Test ratio ≤ 0.05).

Table 3. Few-shot and zero-shot AD results. For each experiment, we report the mean and standard deviation for AUC (in %), obtained by bootstrapping with 10000 repetitions. Best results by column are in **bold**.

Configuration Train/Test ratio	Zero-shot	Few-shot		
	0	0.05	1	223
Ours	65.41\pm0.36	66.02\pm0.02	86.95 \pm 0.23	91.18 \pm 0.19
LDDMM	54.6 \pm 0.36	58.68 \pm 0.02	89.43\pm0.19	95.41\pm0.11
ASM	58.42 \pm 0.36	61.1 \pm 0.02	82.64 \pm 0.26	93.79 \pm 0.14

References

1. Bône, A., Louis, M., Martin, B., Stanley, D.: Deformetrica 4: An open-source software for statistical shape analysis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 3–13. Springer (2018)
2. Cootes, T.F., Taylor, C.J., Graham, J.: Active shape models-their training and application. Computer Vision and Image Understanding **61**(1), 38–50 (1995)
3. Gori, P., Colliot, O., Marrakchi-Kacem, L., Worbe, Y., Poupon, C., Hartmann, A., Ayache, N., Durrleman, S.: A Bayesian Framework for Joint Morphometry of Surface and Curve meshes in Multi-Object Complexes. Medical Image Analysis **35**, 458–474 (2017)
4. Isensee, F., et al.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**(2), 203–211 (2021)
5. Kingma, D.P., Welling, M.: Auto-Encoding Variational Bayes. In: 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada (2014)