

# Supplemental Material for: Is Visual in-Context Learning for Compositional Medical Tasks within Reach?

Simon Reiß<sup>1</sup> ✉ Zdravko Marinov<sup>1</sup> Alexander Jaus<sup>1</sup> Constantin Seibold<sup>2</sup>  
M. Saquib Sarfraz<sup>1,3</sup> Erik Rodner<sup>4</sup> Rainer Stiefelhagen<sup>1</sup>

<sup>1</sup>Karlsruhe Institute of Technology <sup>2</sup>Heidelberg University Hospital <sup>3</sup>Mercedes-Benz Tech Innovation <sup>4</sup>University of Applied Sciences Berlin

✉ [simon.reiss@kit.edu](mailto:simon.reiss@kit.edu)

## Abstract

*In this document, we aim to supply the interested reader with additional side-information on visual in-context learning on compositional medical tasks. This information includes more details on the datasets used for training, the synthetic tasks we trained on, the experimental settings and models we trained and how we evaluated them. We further provide more qualitative outputs and information on used open-source libraries. Finally, we name limitations and discuss the societal impact of our work.*

## 1. Dataset

Our training dataset is based on the MedSAM dataset [33], more specifically, the 38 datasets of which we show examples in Table 1, more details can be found in Table 2. As can be seen, the datasets cover a wide variety of imaging modalities.

We enrich the data with synthetic tasks, which we put into task sequences for compositional medical tasks as we described in the main paper. An example of the individual generative-, transformation- and discriminative tasks we extract from each image-segmentation pair is shown in Tab. 3.

## 2. Information on experiments in Chapter 4. Analysis of task recovery in codebooks

As we outline in the main paper, we identify the limitations that are induced when operating in the token space of learned codebooks. These codebooks entail certain reconstruction errors, which we explored for simple image reconstruction and the reconstruction of segmentation maps on the datasets DAP ATLAS [25], BSDS/CBSD68 [35], HAM10000 [48], Openorganelle (hela-2, hela-3, jurkat-1, macrophage-2, sum-159) [20] and RETOUCH (cirrus, topcon, spectralis) [12]. In Table 4, we show the results

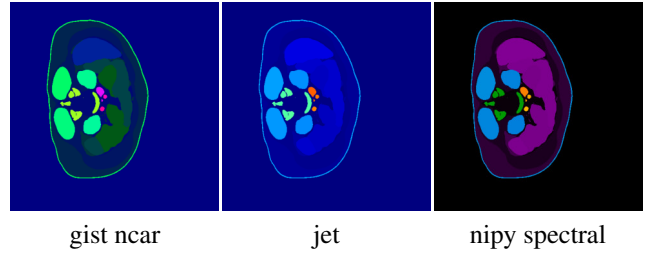


Figure 1. Different colormaps for the ATLAS dataset with which we supply codebooks to indicate the effect of visual prompting.

from the spiderplots in the main paper in numerical form. There, results for auto-encoding images, noisy images and segmentation maps are reported. For the ATLAS dataset with its 142 classes, we briefly investigated the significance of different color schemes when processing segmentation maps. To do this, we utilized color-maps from matplotlib<sup>1</sup>, the names of which are indicated in Table 4. In Figure 1 we show a segmentation map with these different color-maps.

The different VQ-GAN models we test are pre-trained models from a model zoo<sup>2</sup> which is associated to [11, 43]. The models vary in their loss function in training and network architectures which influences their codebook size and latent tensor shape in the quantization layer. For a detailed description of their differences we kindly refer the reader to the explanations by Rombach *et al.* [43].

## 3. Hyperparameters for VQ-GAN and transformer training

In Table 5, we summarize the general hyperparameters of the VQ-GAN models and the transformer models we train.

<sup>1</sup><https://matplotlib.org/stable/users/explain/colors/colormaps.html>

<sup>2</sup><https://github.com/CompVis/latent-diffusion/tree/main?tab=readme-ov-file#pretrained-autoencoding-models>

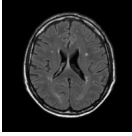



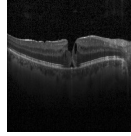
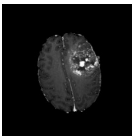
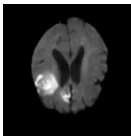
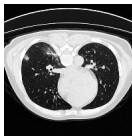
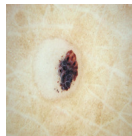

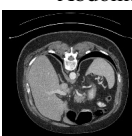

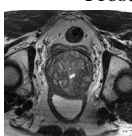
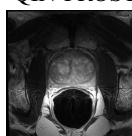
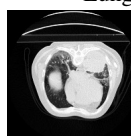



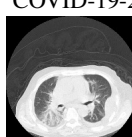


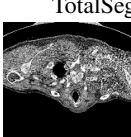
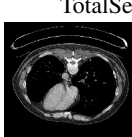

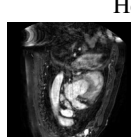


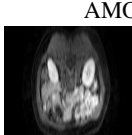
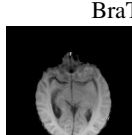
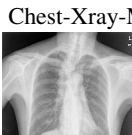
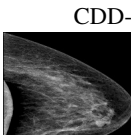




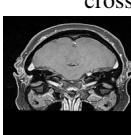
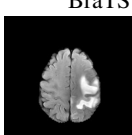
WMH FLAIR 	SpineMR 	ProstateADC 	COVID-19-Radiogr.-Datab. 	Intraretinal-Cystoid-Fluid 
BraTS T1CE 	ISLES2022 DWI 	COVID19-CT-Seg-Bench 	ISIC2018 	m2caiSeg 
AbdomenCT1K 	AMOS 	ProstateT2 	QIN-PROSTATE-Prostate 	LungMasks 
Breast-Ultrasound 	Pneumothorax-Masks 	Kvasir-SEG 	COVID-19-20-CT-SegCh. 	LIDC-IDRI 
LungLesions 	TotalSeg muscles 	TotalSeg organs 	CT-ORG 	Heart 
hc18 	TotalSeg cardiac 	AMOSMR 	BraTS T1 	QIN-PROSTATE-Lesion 
Chest-Xray-Masks-Labels 	CDD-CESM 	TCIA-LCTSC 	ISLES2022 ADC 	COVID-QU-Ex-I.M. Lung 
CholecSeg8k 	crossmoda 	BraTS FLAIR 		

Table 1. Example images and segmentations from 38 datasets in MedSAM [33] which we use as basis for our investigation.

Dataset	Link
AMOSMR [27]	<a href="https://amos22.grand-challenge.org/Dataset/">https://amos22.grand-challenge.org/Dataset/</a>
BraTS T1 [7–10, 37]	<a href="http://braintumorsegmentation.org/">http://braintumorsegmentation.org/</a>
BraTS T1CE [7–10, 37]	<a href="http://braintumorsegmentation.org/">http://braintumorsegmentation.org/</a>
CDD-CESM [28]	<a href="https://www.cancerimagingarchive.net/collection/cdd-cesm/">https://www.cancerimagingarchive.net/collection/cdd-cesm/</a>
Chest-Xray-Masks-Labels [13, 24]	<a href="https://www.kaggle.com/datasets/nikhilpandey360/chest-xray-masks-and-labels">https://www.kaggle.com/datasets/nikhilpandey360/chest-xray-masks-and-labels</a>
COVID-19-20-CT-SegCh. [3, 44]	<a href="https://covid-segmentation.grand-challenge.org/Data/">https://covid-segmentation.grand-challenge.org/Data/</a>
COVID-QU-Ex-I.M. Lung. [14, 17, 41, 46, 47]	<a href="https://www.kaggle.com/datasets/anasmohammedtahir/covidqu">https://www.kaggle.com/datasets/anasmohammedtahir/covidqu</a>
COVID19-CT-Seg-Bench [30]	<a href="https://github.com/JunMall1/COVID-19-CT-Seg-Benchmark">https://github.com/JunMall1/COVID-19-CT-Seg-Benchmark</a>
crossmoda [18]	<a href="https://crossmoda-challenge.ml/">https://crossmoda-challenge.ml/</a>
hc18 [50]	<a href="https://hc18.grand-challenge.org/">https://hc18.grand-challenge.org/</a>
Heart [45]	<a href="http://medicaldecathlon.com/">http://medicaldecathlon.com/</a>
Intraretinal-Cystoid-Fluid [1]	<a href="https://www.kaggle.com/datasets/zeeshanahmed13/intraretinal-cystoid-fluid">https://www.kaggle.com/datasets/zeeshanahmed13/intraretinal-cystoid-fluid</a>
ISLES2022 ADC [21]	<a href="https://www.isles-challenge.org/">https://www.isles-challenge.org/</a>
ISLES2022 DWI [21]	<a href="https://www.isles-challenge.org/">https://www.isles-challenge.org/</a>
Kvasir-SEG [26, 40]	<a href="https://datasets.simula.no/kvasir/">https://datasets.simula.no/kvasir/</a>
LungLesions [4]	<a href="http://medicaldecathlon.com/">http://medicaldecathlon.com/</a>
m2caiSeg [34]	<a href="https://www.kaggle.com/datasets/salmanmaq/m2caiseg">https://www.kaggle.com/datasets/salmanmaq/m2caiseg</a>
Pneumothorax-Masks [53]	<a href="https://www.kaggle.com/datasets/vbookshelf/pneumothorax-chest-xray-images-and-masks">https://www.kaggle.com/datasets/vbookshelf/pneumothorax-chest-xray-images-and-masks</a>
ProstateADC [45]	<a href="http://medicaldecathlon.com">http://medicaldecathlon.com</a>
ProstateT2 [45]	<a href="http://medicaldecathlon.com">http://medicaldecathlon.com</a>
QIN-PROSTATE-Lesion [4, 19]	<a href="http://doi.org/10.7937/K9/TCIA.2018.MR1CKGND">http://doi.org/10.7937/K9/TCIA.2018.MR1CKGND</a>
QIN-PROSTATE-Prostate [4, 19]	<a href="http://doi.org/10.7937/K9/TCIA.2018.MR1CKGND">http://doi.org/10.7937/K9/TCIA.2018.MR1CKGND</a>
SpineMR [54]	<a href="https://www.cg.informatik.uni-siegen.de/en/spine-segmentation-and-analysis">https://www.cg.informatik.uni-siegen.de/en/spine-segmentation-and-analysis</a>
TCIA-LCTSC [52]	<a href="https://www.cancerimagingarchive.net/collection/lctsc/">https://www.cancerimagingarchive.net/collection/lctsc/</a>
WMH flair [29]	<a href="https://wmh.isi.uu.nl/">https://wmh.isi.uu.nl/</a>
BraTS flair [7–10, 37]	<a href="http://braintumorsegmentation.org/">http://braintumorsegmentation.org/</a>
Breast-Ultrasound [2]	<a href="https://www.kaggle.com/datasets/aryashah2k/breast-ultrasound-images-dataset">https://www.kaggle.com/datasets/aryashah2k/breast-ultrasound-images-dataset</a>
COVID-19-Radiogr.-Datab. [14, 41]	<a href="https://www.kaggle.com/datasets/tawsifurrahman/covid19-radiography-database">https://www.kaggle.com/datasets/tawsifurrahman/covid19-radiography-database</a>
ISIC2018 [15, 16, 48]	<a href="https://challenge.isic-archive.com/data/">https://challenge.isic-archive.com/data/</a>
CholecSeg8k [23, 49]	<a href="https://www.kaggle.com/datasets/newslab/cholecseg8k">https://www.kaggle.com/datasets/newslab/cholecseg8k</a>
TotalSeg cardiac [51]	<a href="https://zenodo.org/records/6802614">https://zenodo.org/records/6802614</a>
CT-ORG [42]	<a href="https://www.cancerimagingarchive.net/collection/ct-org/">https://www.cancerimagingarchive.net/collection/ct-org/</a>
TotalSeg organs [51]	<a href="https://zenodo.org/records/6802614">https://zenodo.org/records/6802614</a>
TotalSeg muscles [51]	<a href="https://zenodo.org/records/6802614">https://zenodo.org/records/6802614</a>
LIDC-IDRI [5]	<a href="https://www.cancerimagingarchive.net/collection/lidc-idri/">https://www.cancerimagingarchive.net/collection/lidc-idri/</a>
AMOS [27]	<a href="https://amos22.grand-challenge.org/">https://amos22.grand-challenge.org/</a>
AbdomenCT1K [31, 32]	<a href="https://github.com/JunMall1/AbdomenCT-1K">https://github.com/JunMall1/AbdomenCT-1K</a>
LungMasks [22]	<a href="https://github.com/JoHof/lungmask">https://github.com/JoHof/lungmask</a>

Table 2. Datasets with their citations and links for further information.

Both architectures and the effects of the hyperparameters in source code can be found in the model definitions of either VQ-GAN<sup>3</sup> or the GPT2 transformer<sup>4</sup>.

## 4. Additional qualitative results

### 4.1. VQ-GAN reconstruction

In the main paper, we present quantitative results regarding the upper bound of codebooks on the MedSAM datasets. There, we evaluate a pre-trained VQ-GAN and different fine-tuned variants which are trained on task data and in-domain images. In Figure 2, we show the difference, that our proposed fine-tuning (color remapping augmentation, dataset- and task balancing) can make for representing task data as compared to an ImageNet pre-trained model. It is clearly evident, that the pre-trained reconstructions are not able to capture the semantic content encoded in the segmentation maps, while fine-tuning shows visually coherent re-

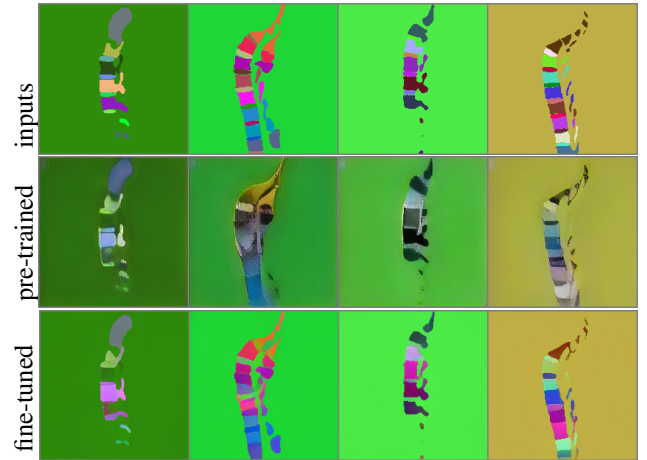


Figure 2. Qualitative examples of auto-encoding segmentation maps with VQ-GAN codebooks. First row shows inputs, second row shows reconstructions with an ImageNet pre-trained VQ-GAN, row three shows a model fine-tuned on task data with color remapping augmentation and dataset + task balancing.

<sup>3</sup><https://github.com/CompVis/taming-transformers/blob/master/taming/models/vqgan.py>

<sup>4</sup><https://github.com/CompVis/taming-transformers/blob/master/taming/modules/transformer/mingpt.py>

sults with only minor inaccuracies in color and shape.

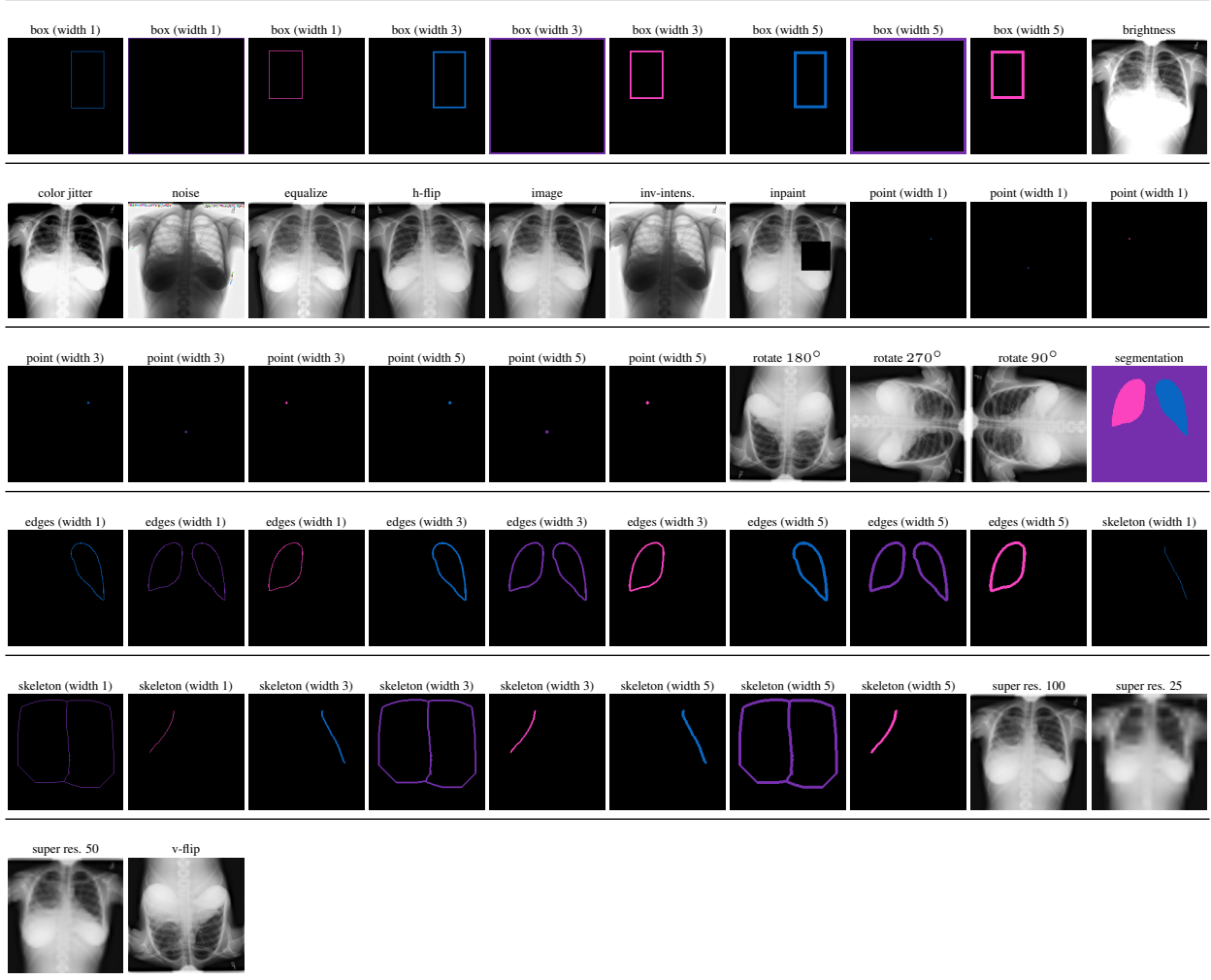


Table 3. Here, we show an example of how each image-segmentation pair is enriched with additional synthetic tasks.

The quantitative evaluation for reconstructing different task-related images (*i.e.*, Table 2 in the main paper) is carried out on 22K test images and segmentation annotations comprised of samples from the 38 datasets we use to train.

#### 4.2. Qualitative compositional predictions

In Figure 3 and Figure 4 we display additional compositional task sequences and associated predictions by the visual in-context transformer. We can observe, that the visual in-context learner, even though the compositional tasks are quite complex and encompass long sequences, can coherently follow the instructions in the context task sequence. Of course, we can also observe inaccuracies, such that certain structures can not be predicted coherently, such as detailed structures in the brain in Figure 3 (top).

Interestingly, when inspecting the image modalities of both Figure 3 and Figure 4, the wide range of modalities, *i.e.*, magnetic resonance imaging, computed tomogra-

phy, ultrasound, and even endoscopy images can be captured nicely.

In some cases, we see, that the visual in-context learner can capture the distribution of the expected output, *e.g.*, the segmentation in the last two columns of Figure 3, in the example at the bottom, but the predicted structures are located at the wrong position. This might be a hint that in case the model is not able to connect the image pattern to the semantic structure, as defined in the context set, it hallucinates a wrong but overall plausible-looking output. Fine, small structures in medical imaging datasets might elevate this problem which also highlights the challenging setting.

#### 5. Limitations

Despite the promising results, the approach has several limitations that need to be addressed in future work. The granularity of what the codebooks can represent limits

Codebook	ATLAS																													
	ATLAS gist near (seg)	ATLAS jet (seg)	ATLAS nipy spectral (seg)	BSDS500 (img)	CBS08 (img)	CBS08 (img noisy5)	CBS08 (img noisy10)	CBS08 (img noisy15)	CBS08 (img noisy25)	CBS08 (img noisy35)	CBS08 (img noisy50)	HAM10000 (img)	HAM10000 (seg)	hela-2 (img)	hela-2 (seg)	hela-3 (img)	hela-3 (seg)	jurkat-1 (img)	jurkat-1 (seg)	macrophage-2 (img)	macrophage-2 (seg)	sum159-1 (img)	sum159-1 (seg)	retouch cirrus (img)	retouch cirrus (seg)	retouch spectralis (img)	retouch spectralis (seg)	retouch topcon (img)	retouch topcon (seg)	
	IoU↑	IoU↑	IoU↑	MAE↓	PSNR↑	RMSE↓	RMSE↓	RMSE↓	RMSE↓	RMSE↓	RMSE↓	MAE↓	IoU↑	MAE↓	IoU↑	MAE↓	IoU↑	MAE↓	IoU↑	MAE↓	IoU↑	MAE↓	IoU↑	MAE↓	IoU↑	MAE↓	IoU↑	MAE↓	IoU↑	MAE↓
VQ-GAN f4	0.40	<b>0.29</b>	0.38	0.03	27.39	<b>0.05</b>	<b>0.06</b>	<b>0.07</b>	<b>0.09</b>	<b>0.11</b>	<b>0.14</b>	<b>0.01</b>	1.00	0.01	0.71	<b>0.04</b>	0.81	<b>0.00</b>	<b>0.83</b>	<b>0.00</b>	<b>0.82</b>	<b>0.01</b>	<b>0.96</b>	0.03	0.98	<b>0.01</b>	0.99	<b>0.02</b>	0.98	
VQ-GAN f8	0.30	0.19	0.31	0.05	22.48	0.08	0.09	0.09	0.11	0.14	0.17	0.02	1.00	0.02	0.68	0.07	0.80	0.01	0.81	0.01	0.80	0.01	0.95	0.05	0.98	0.02	0.98	0.03	0.98	
VQ-GAN f8-n256	0.22	0.23	0.28	0.05	21.81	0.09	0.09	0.10	0.12	0.15	0.18	0.02	1.00	0.02	0.65	0.08	0.78	0.01	0.77	0.01	0.78	0.01	0.94	0.04	0.97	0.03	0.97	0.03	0.97	
VQ-GAN f16	0.17	0.18	0.22	0.06	20.50	0.10	0.11	0.11	0.13	0.15	0.18	0.02	0.99	0.02	0.61	0.09	0.74	0.01	0.75	0.01	0.73	0.01	0.94	0.05	0.95	0.03	0.95	0.04	0.96	
VQ-GAN f4-kl	<b>0.44</b>	0.27	<b>0.43</b>	<b>0.03</b>	<b>27.46</b>	0.05	0.06	0.07	0.09	0.12	0.15	0.01	<b>1.00</b>	<b>0.01</b>	<b>0.71</b>	0.04	0.81	0.01	0.82	0.01	0.80	0.01	0.96	<b>0.03</b>	<b>0.99</b>	0.01	<b>0.99</b>	0.02	<b>0.99</b>	
VQ-GAN f8-kl	0.33	0.28	0.35	0.04	23.65	0.07	0.08	0.09	0.11	0.13	0.16	0.02	1.00	0.02	0.69	0.07	<b>0.81</b>	0.01	0.82	0.01	0.80	0.01	0.95	0.04	0.97	0.02	0.98	0.03	0.98	
VQ-GAN f16-kl	0.34	0.28	0.34	0.04	23.51	0.07	0.08	0.09	0.11	0.13	0.16	0.02	1.00	0.02	0.68	0.07	0.80	0.01	0.80	0.01	0.79	0.01	0.96	0.04	0.98	0.02	0.98	0.03	0.98	
VQ-GAN f32-kl	0.27	0.22	0.28	0.05	21.80	0.09	0.09	0.10	0.12	0.14	0.17	0.02	1.00	0.02	0.63	0.08	0.77	0.01	0.77	0.01	0.76	0.01	0.95	0.04	0.97	0.03	0.97	0.03	0.97	

Table 4. Preliminary evaluation of the ImageNet pre-trained codebooks from Figure 3 in the main paper represented in numerical form. The different rows show different pretrained Codebooks, columns indicate different datasets where it is indicated in brackets whether the images of the datasets are autoencoded or the segmentation maps are autoencoded to determine the upper performance bounds with respective codebooks.

Hyperparameter	VQ-GAN	Transformer
learning rate	$4.5e - 06$	$4.5e - 06$
batch size	96	4
codebook size	16384	$16384 + 2$
codebook enc. resolutions	1, 1, 2, 2, 4	N/A
<sup>†</sup> residual blocks per level	2	N/A
block size	N/A	4, 500
# encoder layers	N/A	6
# attention heads	N/A	16
embedding dimension	N/A	1, 408
dropout ratio (embedding)	N/A	0.1
dropout ratio (residual)	N/A	0.1
dropout ratio (attention)	N/A	0.1
hardware	4×NVIDIA A100 40GB	4×NVIDIA A100 40GB

Table 5. Overview of the hyperparameters for both the VQ-GAN models and the GPT2 transformers we train.

their ability to recover fine-grained structures and, while color remapping augmentation improves generalization, the model remains sensitive to color variations where classes are encoded in very similar colors (*e.g.*, different shades of blue encoding different classes).

Ensuring consistency across intermediate outputs of the visual in-context learner is a remaining challenge as well. The model currently may produce misaligned sub-task predictions, affecting the overall coherence in the output sequence. This also ties into the qualitative observation that with longer sequences the intermediate outputs degrade successively (*e.g.*, Figure 3, second row where late in the sequence, the predicted brain scan is tinted blue).

The synthetic task generation pipeline, although effective, may not fully capture the complexity and diversity of real-world tasks which might impact the model’s gen-

eralization capabilities. Exploring the effects of adapting this task generation pipeline from a data-centric perspective would be beneficial.

Balancing training data between imaging data and task outputs is crucial but challenging, especially with diverse datasets with different amounts of samples and the different generative-, transformation- and discriminative tasks.

Evaluation metrics such as IoU, F-1 Score, MAE, RMSE, and PSNR, while useful, may not fully capture the nuances of complex task sequences, necessitating more comprehensive evaluation frameworks. Here, efforts towards designing a compositional medical task benchmark are needed which encompasses a wide variety of compositional task prompts and associated evaluation schemes.

Finally, while the model shows promising results on complex, compositional task sequences, there is still much room for improvement when moving to out-of-domain sequences, *i.e.*, when moving beyond training datasets. One elevating factor for the presented approach could lie in model- and data scale, as visual in-context learners that were trained on larger datasets, with higher parameter counts [6] exhibited strong out-of-domain capabilities.

Addressing these limitations will be crucial for advancing visual in-context learning and enabling more robust, adaptable models for a wide range of applications. Future work should focus on enhancing codebooks, improving training strategies, expanding the synthetic task pipeline, and developing comprehensive evaluation frameworks.

## 6. Utilized open source code and libraries

We heavily build on the two codebases<sup>5 6</sup> and their open-source models. The GPT-2 implementation is based on the repository of minGPT<sup>7</sup>. Further, we make heavy use of li-

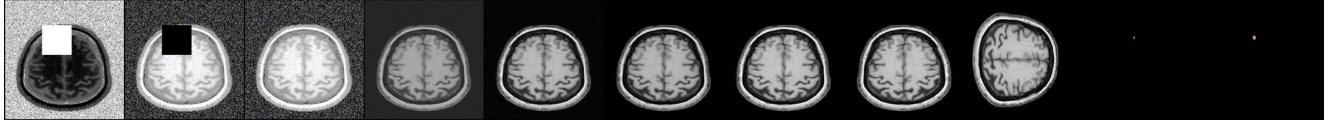
<sup>5</sup><https://github.com/CompVis/taming-transformers>

<sup>6</sup><https://github.com/CompVis/latent-diffusion>

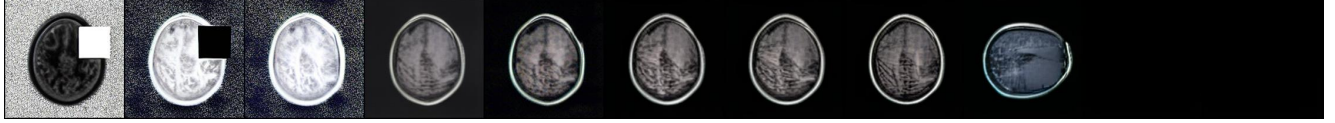
<sup>7</sup><https://github.com/karpathy/minGPT/>



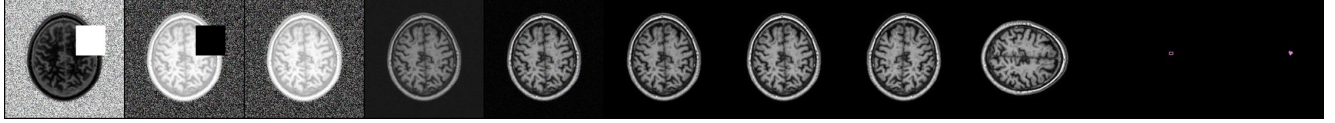
Context compositional task sequence



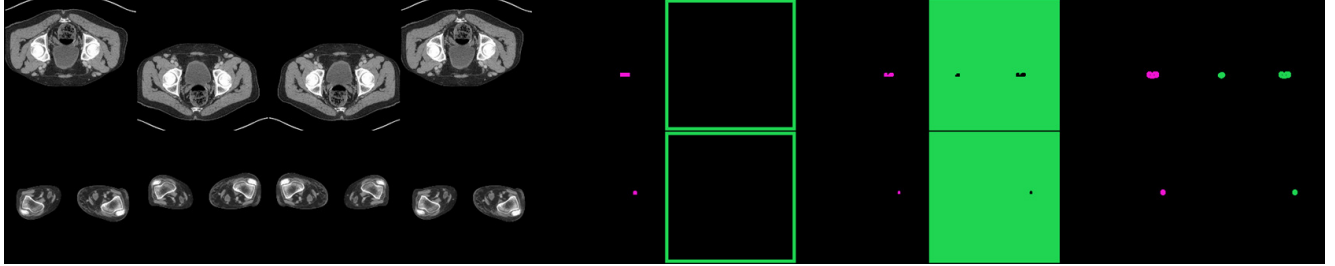
Query Predicted sequential output



Query Ground-truth compositional task sequence



Context compositional task sequence



Query Predicted sequential output



Query Ground-truth compositional task sequence

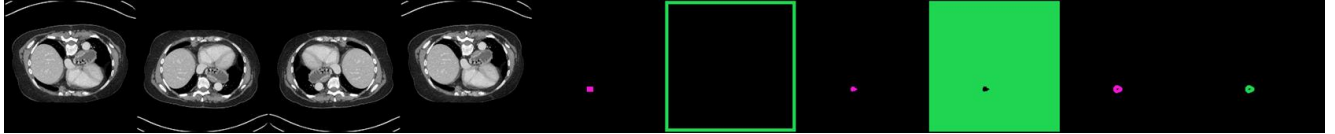


Figure 3. Additional compositional contexts, predicted task sequences and the associated ground-truth task sequences.

libraries such as Pytorch [38], Pandas [36] and sklearn [39].

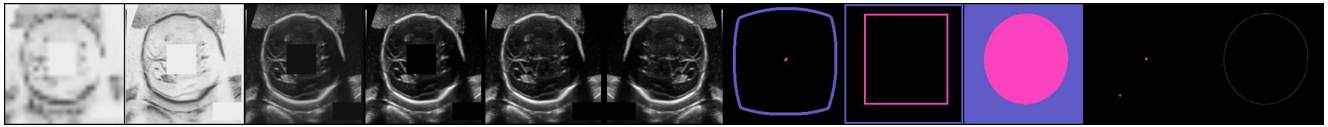
## 7. Impact statement

This paper presents work whose goal is to explore machine learning techniques for solving compositional visual tasks from the field of medical image analysis. Specifically, our goal is to enable users to specify vision-task pipelines without the requirement of programming knowledge or model re-training. This goal inherently means, that a broader audience could be able to use the developed models, which may include users with malicious intent. As the current models are not able to produce results that are on a level to be used

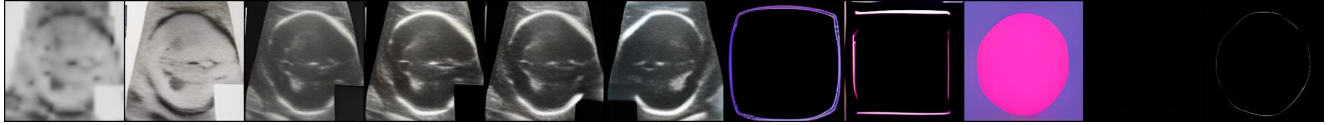
in medical practice and mainly serve to advance research, the factor of misuse may be of relatively low concern.

We utilize a broad set of medical datasets for this exploration and train models on it. As such, these models may reflect the biases within these datasets, such as gender-, age- or ethnicity imbalance.

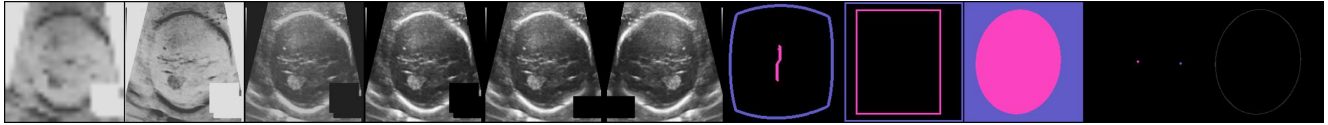
Context compositional task sequence



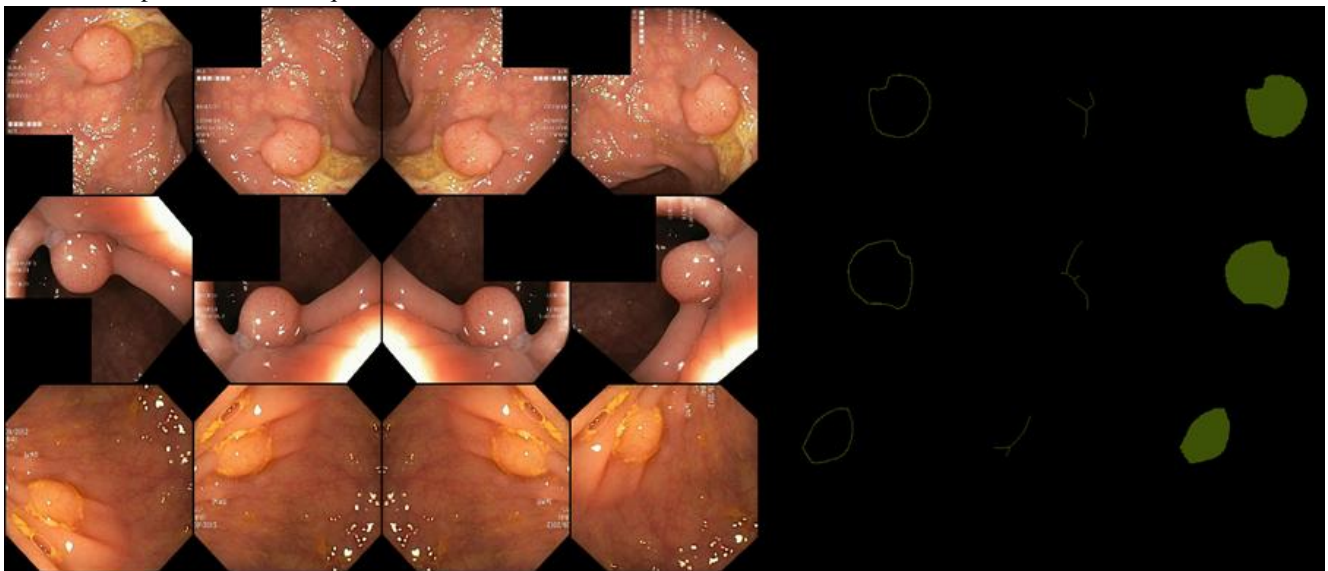
Query Predicted sequential output



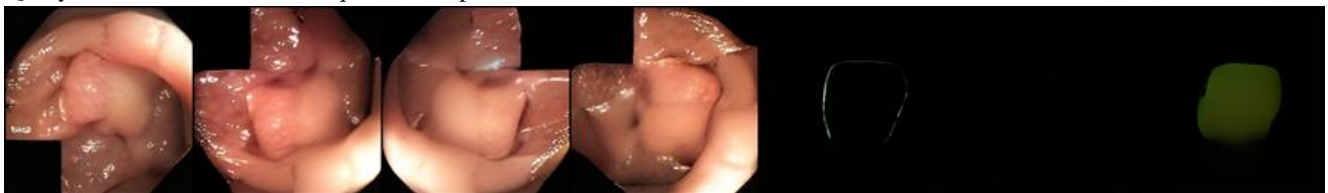
Query Ground-truth compositional task sequence



Context compositional task sequence



Query Predicted sequential output



Query Ground-truth compositional task sequence



Figure 4. Additional compositional contexts, predicted task sequences and the associated ground-truth task sequences.

## References

- [1] Zeeshan Ahmed, Shahbaz Qamar Panhwar, Attiya Baqai, Fahim Aziz Umrani, Munawar Ahmed, and Arbaaz Khan. Deep learning based automated detection of intraretinal cystoid fluid. *International Journal of Imaging Systems and Technology*, 32(3):902–917, 2022. 3
- [2] Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. Dataset of breast ultrasound images. *Data in brief*, 28:104863, 2020. 3
- [3] P. An, S. Xu, S. A. Harmon, E. B. Turkbey, T. H. Sanford, A. Amalou, M. Kassim, N. Varble, M. Blain, V. Anderson, G. Patella, F. and Carrafiello, B. T. Turkbey, and B. J. Wood. Ct images in covid-19 [data set]., 2020. 3
- [4] Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, et al. The medical segmentation decathlon. *Nature communications*, 13(1):4128, 2022. 3
- [5] SG Armato III, G McLennan, L Bidaut, MF McNitt-Gray, CR Meyer, AP Reeves, B Zhao, DR Aberle, CI Henschke, EA Hoffman, et al. Data from lidc-idri [data set]. the cancer imaging archive, 2015. 3
- [6] Yutong Bai, Xinyang Geng, Karttikeya Mangalam, Amir Bar, Alan Yuille, Trevor Darrell, Jitendra Malik, and Alexei A Efros. Sequential modeling enables scalable learning for large vision models. *arXiv preprint arXiv:2312.00785*, 2023. 5
- [7] Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin Kirby, John Freymann, Keyvan Farahani, and Christos Davatzikos. Segmentation labels and radiomic features for the pre-operative scans of the tcga-gbm collection (2017). DOI: <https://doi.org/10.7937/K,9,2017.3>
- [8] Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, JS Kirby, JB Freymann, Keyvan Farahani, and Christos Davatzikos. Segmentation labels and radiomic features for the pre-operative scans of the tcga-lgg collection [data set]. the cancer imaging archive, 2017.
- [9] Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S Kirby, John B Freymann, Keyvan Farahani, and Christos Davatzikos. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Scientific data*, 4(1): 1–13, 2017.
- [10] Spyridon Bakas, Mauricio Reyes, Andras Jakab, Stefan Bauer, Markus Rempfler, Alessandro Crimi, Russell Takeshi Shinohara, Christoph Berger, Sung Min Ha, Martin Rozycki, et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. *arXiv preprint arXiv:1811.02629*, 2018. 3
- [11] Andreas Blattmann, Robin Rombach, Kaan Oktay, and Björn Ommer. Retrieval-augmented diffusion models, 2022. 1
- [12] Hrvoje Bogunović, Freerk Venhuizen, Sophie Klimscha, Stefanos Apostolopoulos, Alireza Bab-Hadiashar, Ulas Bagci, Mirza Faisal Beg, Loza Bekalo, Qiang Chen, Carlos Ciller, et al. Retouch: The retinal oct fluid detection and segmentation benchmark and challenge. *IEEE transactions on medical imaging*, 38(8):1858–1874, 2019. 1
- [13] Sema Candemir, Stefan Jaeger, Kannappan Palaniappan, Jonathan P Musco, Rahul K Singh, Zhiyun Xue, Alexandros Karargyris, Sameer Antani, George Thoma, and Clement J McDonald. Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration. *IEEE transactions on medical imaging*, 33(2):577–590, 2013. 3
- [14] Muhammad EH Chowdhury, Tawsifur Rahman, Amith Khandakar, Rashid Mazhar, Muhammad Abdul Kadir, Zaid Bin Mahbub, Khandakar Reajul Islam, Muhammad Salman Khan, Atif Iqbal, Nasser Al Emadi, et al. Can ai help in screening viral and covid-19 pneumonia? *Ieee Access*, 8:132665–132676, 2020. 3
- [15] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*, 2019. 3
- [16] Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pages 168–172. IEEE, 2018. 3
- [17] Aysen Degerli, Mete Ahishali, Mehmet Yamac, Serkan Kiranyaz, Muhammad EH Chowdhury, Khalid Hameed, Tahir Hamid, Rashid Mazhar, and Moncef Gabbouj. Covid-19 infection map generation and detection from chest x-ray images. *Health information science and systems*, 9(1):15, 2021. 3
- [18] Reuben Dorent, Aaron Kujawa, Marina Ivory, Spyridon Bakas, Nicola Rieke, Samuel Joutard, Ben Glocker, Jorge Cardoso, Marc Modat, Kayhan Batmanghelich, et al. Crossmoda 2021 challenge: Benchmark of cross-modality domain adaptation techniques for vestibular schwannoma and cochlea segmentation. *Medical Image Analysis*, 83:102628, 2023. 3
- [19] Andriy Fedorov, Michael Schmier, David Clunie, Christian Herz, Steve Pieper, Ron Kikinis, Clare Tempany, and Fiona Fennessy. An annotated test-retest collection of prostate multiparametric mri. *Scientific data*, 5(1):1–13, 2018. 3
- [20] Larissa Heinrich, Davis Bennett, David Ackerman, Woohyun Park, John Bogovic, Nils Eckstein, Alyson Petruncio, Jody Clements, Song Pang, C Shan Xu, et al. Whole-cell organelle segmentation in volume electron microscopy. *Nature*, 599(7883):141–146, 2021. 1
- [21] Moritz R Hernandez Petzsche, Ezequiel de la Rosa, Uta Hanning, Roland Wiest, Waldo Valenzuela, Mauricio Reyes, Maria Meyer, Sook-Lei Liew, Florian Kofler, Ivan Ezhov, et al. Isles 2022: A multi-center magnetic resonance imag-



- ing stroke lesion segmentation dataset. *Scientific data*, 9(1): 762, 2022. 3
- [22] Johannes Hofmanninger, Florian Prayer, Jeanny Pan, Sebastian Röhrich, Helmut Prosch, and Georg Langs. Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem. *European radiology experimental*, 4:1–13, 2020. 3
- [23] W-Y Hong, C-L Kao, Y-H Kuo, J-R Wang, W-L Chang, and C-S Shih. Cholecseg8k: a semantic segmentation dataset for laparoscopic cholecystectomy based on cholec80. *arXiv preprint arXiv:2012.12453*, 2020. 3
- [24] Stefan Jaeger, Alexandros Karargyris, Sema Candemir, Les Folio, Jenifer Siegelman, Fiona Callaghan, Zhiyun Xue, Kannappan Palaniappan, Rahul K Singh, Sameer Antani, et al. Automatic tuberculosis screening using chest radiographs. *IEEE transactions on medical imaging*, 33(2):233–245, 2013. 3
- [25] Alexander Jaus, Constantin Seibold, Kelsey Hermann, Negar Shahamiri, Alexandra Walter, Kristina Giske, Johannes Haubold, Jens Kleesiek, and Rainer Stiefelhagen. Towards unifying anatomy segmentation: Automated generation of a full-body ct dataset. In *2024 IEEE International Conference on Image Processing (ICIP)*, pages 41–47. IEEE, 2024. 1
- [26] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas De Lange, Dag Johansen, and Håvard D Johansen. Kvasir-seg: A segmented polyp dataset. In *International conference on multimedia modeling*, pages 451–462. Springer, 2019. 3
- [27] Yuanfeng Ji, Haotian Bai, Jie Yang, Chongjian Ge, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhang, Wanling Ma, Xiang Wan, et al. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. *arXiv preprint arXiv:2206.08023*, 2022. 3
- [28] R Khaled et al. Categorized digital database for low energy and subtracted contrast enhanced spectral mammography images. *The Cancer Imaging Archive*, 2021. 3
- [29] Hugo J Kuijf, J Matthijs Biesbroek, Jeroen De Bresser, Rutger Heinen, Simon Andermatt, Mariana Bento, Matt Berseth, Mikhail Belyaev, M Jorge Cardoso, Adria Casamitjana, et al. Standardized assessment of automatic segmentation of white matter hyperintensities and results of the wmh segmentation challenge. *IEEE transactions on medical imaging*, 38(11): 2556–2568, 2019. 3
- [30] Jun Ma, Yixin Wang, Xingle An, Cheng Ge, Ziqi Yu, Jianan Chen, Qiongjie Zhu, Guoqiang Dong, Jian He, Zhiqiang He, et al. Toward data-efficient learning: A benchmark for covid-19 ct lung and infection segmentation. *Medical physics*, 48(3):1197–1210, 2021. 3
- [31] Jun Ma, Yao Zhang, Song Gu, Xingle An, Zhihe Wang, Cheng Ge, Congcong Wang, Fan Zhang, Yu Wang, Yinan Xu, et al. Fast and low-gpu-memory abdomen ct organ segmentation: the flare challenge. *Medical Image Analysis*, 82: 102616, 2022. 3
- [32] Jun Ma, Yao Zhang, Song Gu, Cheng Zhu, Cheng Ge, Yichi Zhang, Xingle An, Congcong Wang, Qiyuan Wang, Xin Liu, Shucheng Cao, Qi Zhang, Shangqing Liu, Yunpeng Wang, Yuhui Li, Jian He, and Xiaoping Yang. Abdomenct-1k: Is abdominal organ segmentation a solved problem? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6695–6714, 2022. 3
- [33] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024. 1, 2
- [34] Salman Maqbool, Aqsa Riaz, Hasan Sajid, and Osman Hasan. m2caiseg: Semantic segmentation of laparoscopic images using convolutional neural networks. *arXiv preprint arXiv:2008.10134*, 2020. 3
- [35] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proc. 8th Int’l Conf. Computer Vision*, pages 416–423, 2001. 1
- [36] Wes McKinney et al. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, pages 51–56. Austin, TX, 2010. 6
- [37] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993–2024, 2014. 3
- [38] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. 6
- [39] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. 6
- [40] Konstantin Pogorelov, Kristin Ranheim Randel, Carsten Griwodz, Sigrun Losada Eskeland, Thomas de Lange, Dag Johansen, Concetto Spampinato, Duc-Tien Dang-Nguyen, Mathias Lux, Peter Thelin Schmidt, et al. Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection. In *Proceedings of the 8th ACM on Multimedia Systems Conference*, pages 164–169, 2017. 3
- [41] Tawsifur Rahman, Amith Khandakar, Yazan Qiblawey, Anas Tahir, Serkan Kiranyaz, Saad Bin Abul Kashem, Mohammad Tariqul Islam, Somaya Al Maadeed, Susu M Zughaier, Muhammad Salman Khan, et al. Exploring the effect of image enhancement techniques on covid-19 detection using chest x-ray images. *Computers in biology and medicine*, 132: 104319, 2021. 3
- [42] Blaine Rister, Darvin Yi, Kaushik Shivakumar, Tomomi Nobashi, and Daniel L Rubin. Ct-org, a new dataset for multiple organ segmentation in computed tomography. *Scientific Data*, 7(1):381, 2020. 3

- [43] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 1
- [44] Holger R Roth, Ziyue Xu, Carlos Tor-Díez, Ramon Sanchez Jacob, Jonathan Zember, Jose Molto, Wenqi Li, Sheng Xu, Baris Turkbey, Evrim Turkbey, et al. Rapid artificial intelligence solutions in a pandemic—the covid-19-20 lung ct lesion segmentation challenge. *Medical image analysis*, 82: 102605, 2022. 3
- [45] Amber L Simpson, Michela Antonelli, Spyridon Bakas, Michel Bilello, Keyvan Farahani, Bram Van Ginneken, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, et al. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv preprint arXiv:1902.09063*, 2019. 3
- [46] Anas M Tahir, Muhammad EH Chowdhury, Amith Khandakar, Tawsifur Rahman, Yazan Qiblawey, Uzair Khurshid, Serkan Kiranyaz, Nabil Ibtehaz, M Sohel Rahman, Somaya Al-Maadeed, et al. Covid-19 infection localization and severity grading from chest x-ray images. *Computers in biology and medicine*, 139:105002, 2021. 3
- [47] Anas M. Tahir, Muhammad E. H. Chowdhury, Yazan Qiblawey, Amith Khandakar, Tawsifur Rahman, Serkan Kiranyaz, Uzair Khurshid, Nabil Ibtehaz, Sakib Mahmud, and Maymouna Ezeddin. Covid-qu-ex dataset, 2022. 3
- [48] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):1–9, 2018. 1, 3
- [49] Andru P Twinanda, Sherif Shehata, Didier Mutter, Jacques Marescaux, Michel De Mathelin, and Nicolas Padoy. Endonet: a deep architecture for recognition tasks on laparoscopic videos. *IEEE transactions on medical imaging*, 36(1):86–97, 2016. 3
- [50] Thomas LA van den Heuvel, Dagmar de Bruijn, Chris L de Korte, and Bram van Ginneken. Automated measurement of fetal head circumference using 2d ultrasound images. *PloS one*, 13(8):e0200412, 2018. 3
- [51] Jakob Wasserthal, Hanns-Christian Breit, Manfred T Meyer, Maurice Pradella, Daniel Hinck, Alexander W Sauter, Tobias Heye, Daniel T Boll, Joshy Cyriac, Shan Yang, et al. Totalsegmentator: robust segmentation of 104 anatomic structures in ct images. *Radiology: Artificial Intelligence*, 5(5): e230024, 2023. 3
- [52] J. Yang, G. Sharp, H. Veeraraghavan, W. Van Elmpt, A. Dekker, T. Lustberg, and M. Gooding. Data from lung ct segmentation challenge (lctsc) (version 3) [data set]., 2017. 3
- [53] Anna Zawacki, Carol Wu, George Shih, Julia Elliott, Mikhail Fomitchev, Mohannad Hussain, ParasLakhani, Phil Culliton, and Shunxing Bao. Siim-acr pneumothorax segmentation. <https://kaggle.com/competitions/siim-acr-pneumothorax-segmentation>, 2019. Kaggle. 3
- [54] Dženan Zukić, Aleš Vlasák, Jan Egger, Daniel Hořínek, Christopher Nimsky, and Andreas Kolb. Robust detection and segmentation for diagnosis of vertebral diseases using routine mr images. In *Computer Graphics Forum*, pages 190–204. Wiley Online Library, 2014. 3