Organ-DETR: Organ Detection via Transformers (Datasets and Code Documentation)

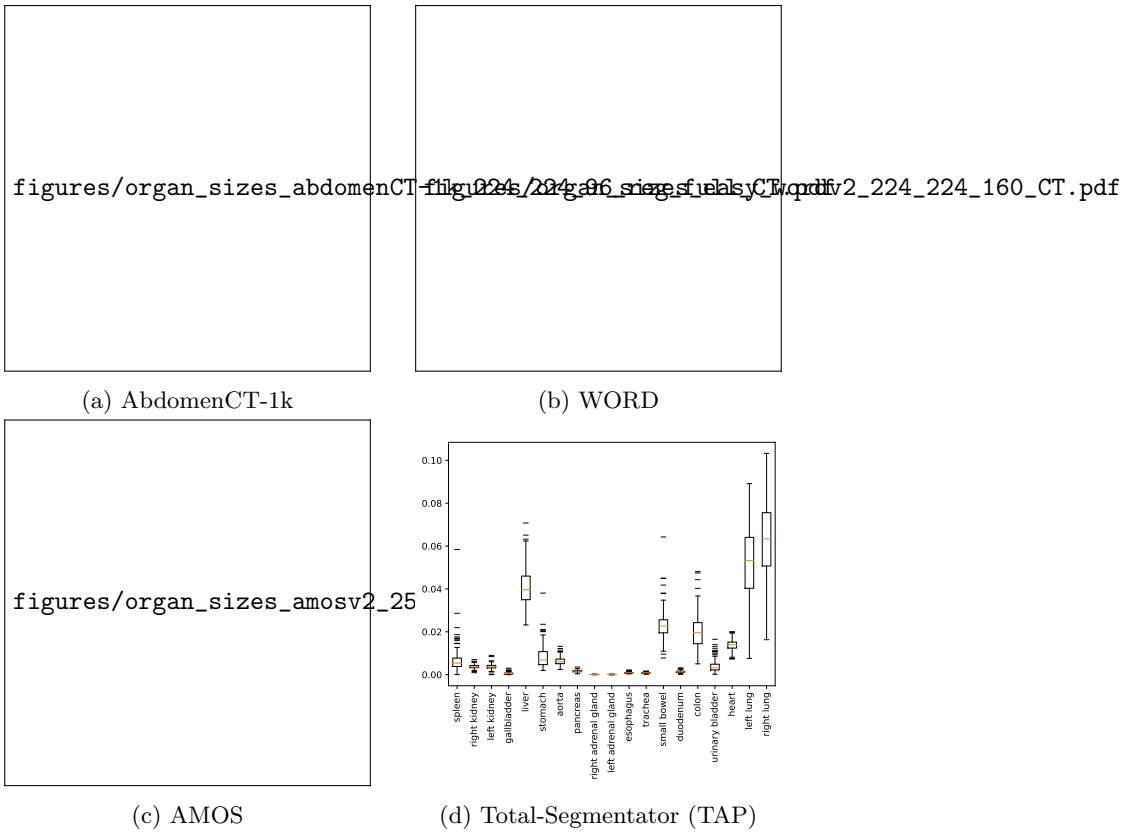
[1]Lab for AI in Medical Imaging (AI-Med),
Department of Radiology, Technical University of Munich (TUM),
[2]Munich Center for Machine Learning (MCML),
Munich, Germany

1 Datasets and data preparation

Organ-DETR is evaluated on five publicly available CT datasets. The datasets with preprocessing and data augmentation are detailed below. All datasets provide segmentation labels as ground truth. During training, axis-aligned bounding boxes are extracted from the segmentation maps for each class to generate ground truth bounding boxes. An overview of considered organs and their relative sizes in the image volume can be found in Fig. 1.

1.1 AbdomenCT-1k

AbdomenCT-1k [1] contains 1112 abdominal CT images from different medical centers. For each image, four organs have been labeled with voxel-wise segmentation labels. The images have an axial pixel resolution of 512 $\times$ 512. The slice thickness is between 1.25 and 5 mm. Data was resampled to an isotropic voxel spacing of 2 mm along each axis. All images have been registered to the first scan in the dataset. A subset containing 160 samples was also created for development purposes. The original segmentation labels cover the liver, kidney, spleen, and pancreas. Since bounding boxes are extracted based on the segmentation map, another label was introduced to distinguish the left and right kidneys. Therefore, a script was developed that determines the centers of both kidneys, separates both kidneys via a sagittal plane, and relabels the left kidney. The body's orientation was ensured to be equal in all images using registered data. The orientation was also confirmed by the metadata in the NIfTI files. Pre-processing for the AbdomenCT-1K dataset is straightforward, given its field-of-views (FOVs) that exclusively cover the abdomen. In this procedure, all 975 CT samples underwent orientation standardization to Right-Anterior-Superior (RAS), followed by cropping to include labeled regions with a two-pixel margin, and finally resizing to the specified target dimensions of (224, 224, 96). These essential preprocessing steps were consistently applied across



(a) AbdomenCT-1k (b) WORD

(c) AMOS (d) Total-Segmentator (TAP)

Figure 1: Relative volumes of segmentation maps from different datasets.

Table 1: List of organs in the preprocessed CT datasets

Dataset	List of Organs
AbdomenCT-1K	pancreas, left kidney, right kidney, spleen, liver
WORD	pancreas, duodenum, left kidney, right kidney, spleen, urinary bladder, liver, stomach, small bowel, colon (merged with rectum)
Total-Segmentator	gallbladder, pancreas, esophagus, left adrenal gland, right adrenal gland, trachea, urinary bladder, left kidney, right kidney, spleen, aorta, duodenum, liver, small bowel, colon, stomach, heart, left lung, right lung
AMOS	esophagus, left adrenal gland, right adrenal gland, prostate/uterus, left kidney, right kidney, spleen, pancreas, gallbladder, aorta, postcava, duodenum, urinary bladder
VerSe	vertebrae {C1–C7, T1–T12, L1–L5}

all datasets, ensuring uniformity and compatibility in the dataset preparation process.

1.2 AMOS

AMOS (Abdominal Multi-Organ Segmentation) [2] contains 500 CT images from different medical centers and imaging devices. The segmentation labels cover 15 organs (Table 1). Every CT image is composed of 67 to 369 slices. 385 images have an axial in-plane resolution of 512 × 512 pixels and 115 have a resolution of 768 × 768. The median voxel spacing is 0.67 mm × 0.67 mm × 5.0 mm (min: 0.45 mm × 0.45 mm × 1.25 mm, max: 1.07 mm × 1.07 mm × 5.0 mm). Only the training and validation subsets have been used to create training, validation, and test datasets because the labels for the original test dataset are not publicly available. The liver, stomach, spleen, urinary bladder, and prostate/uterus were considered during the organ cropping process. The transformed scans checked their boundary voxels to confirm the presence of the specified organs within the cropping region. If the margin could not be applied and there was still an organ in the boundary layer of the scan, the scan was skipped. This was done to prevent cropped boundary organs from being in the dataset. Along the preprocessing steps, these boundary organs were defined in [3] for tests. The same set of organs and preprocessing steps were used for this dataset to keep comparability.

1.3 Total-Segmentator

Total-Segmentator [4] contains 1204 CT images from different medical centers, covering a variety of field-of-views. Segmentation labels for 104 anatomical structures, such as bones, muscles, organs, and vessels, are provided. Every CT image is composed of 77 to 486 slices. The in-plane axial pixel resolution is varying. Each volume has a 1.5 mm isotropic spacing. A provided metadata table contains information about field-of-view categories. To evaluate the object detector, a Thorax-Abdomen-Pelvis (TAP) subset of 19 organs (Table 1) has been defined based on the "ct thorax-abdomen-pelvis" and "ct neck-thorax-abdomen-pelvis" categories. The heart and lungs contained different labels for parts of the organs, which were merged into labels "heart", "left lung", and "right lung". The image cropping process involved selecting specific organs, which included the lungs, liver, stomach, spleen, colon, and urinary bladder. Like the AMOS dataset, a designated set of boundary organs was employed to identify organs that might have unintentionally been cropped along the image edges. The segmentation map was relabeled for a defined set of 19 organs (outlined in Table 1) since the original dataset defines 104 anatomical structures.

1.4 VerSe

VerSe (Vertebrae Segmentation) [5–7] contains 374 CT images. All images stem from CT scanners from four different manufacturers. Furthermore, the images cover a variety of field-of-views and abnormalities like fractures, metallic implants, or foreign materials. Segmentation labels are provided for 26 vertebrae. Only 24 classes have been considered for the evaluation of the object detector. L6 and T13 have been removed since very few samples contain these classes. After the initial data split, they were missing entirely in the validation and test dataset. Every CT image is composed of 34 to 2023 slices, which have a resolution of [103, 144] to [960, 2048]. The in-plane spacing reaches from 0.195 mm $\times$ 0.195 mm to 1.675 mm $\times$ 1.675 mm. Slicing thickness varies from 0.4 mm to 5.0 mm. Although this dataset does not contain organs, it is used because it poses as a challenging object detection dataset for CT images. The VerSe dataset contains varying FOVs, so its CT scans' preprocessing differs from the other datasets' preprocessing. A fixed FOV cropping method proved inadequate because of the substantial FOV differences between images. As an alternative, the initial step involved cropping CT scans around any labels with a margin of three. Labels located at the image boundaries were excluded from this process. To ensure uniformity, all scans were resampled to achieve an isotropic spacing of $3mm$. Additionally, the scans were padded to reach a final size of (64,64,256). Two scans, however, posed a challenge as they exceeded the intended target size, rendering the standard preprocessing approach ineffective for them. Consequently, the pre-processed VerSe dataset comprises a total of 372 CT samples.

1.5 WORD

WORD (Whole abdominal ORgan Dataset) [8] contains 150 CT images captured by the same medical center and imaging device. The segmentation labels cover 16 anatomical structures. Every CT image is composed of 159 to 330 slices, each with a resolution of 512 $\times$ 512 pixels. The axial in-plane spacing is 0.976 mm $\times$ 0.976 mm, and the spacing between slices ranges from 2.5 mm to 3.0 mm. For organ detection, the femur heads were excluded. During the preprocessing of WORD, the CT scans were cropped using specific organs, namely the colon, small bowel, spleen, stomach, urinary bladder, and rectum. All these organs, except the liver, were considered when conducting the boundary check. The decision to exclude the liver from the boundary check was based on the observation that only a few voxels of the liver typically touched the image boundary.

1.6 Augmentation setting

CT scans were normalized to fall within the range [0, 1]. This was achieved by scaling the voxel values based on the 0.5 and 99.5 percentiles of the non-background voxels within the input scan. All methods have undergone an identical series of augmentations using MONAI, applied in the following order (Details can be found in the code):

- Random rotation: Scans are randomly rotated by *RandRotated* within a range of -10 to 10 degrees.
- Random scaling: Scans were randomly re-scaled by *RandZoomd*, with a scale factor in the range of 0.9 to 1.1.
- Random translation: Scans were randomly translated by *RandAffined*, with a random shift in -10% to 10% of the input size.
- Random intensity scaling: Scans' intensity values were scaled by *RandScaleIntensityd*, with a random scale in the range of 0% to 10%.
- Random intensity shifting: Scans' intensity values were shifted by *RandShiftIntensityd*, with a random shift from 0% to 10%.

2 Code Details and Parameters setting

2.1 Matchers' setting

Table 2 presents the settings for various matching techniques. The reported settings represent the optimal results achieved for each method; thus, the nominal values are provided here. As discussed in the paper, η in the range [1,5] is the favorable interval for DQM in Organ-DETR.

Table 2: Parameter setting of different matching techniques for each dataset

Method	$\#M$	Hyperparameters
	AbdomenCT-1K	
Hungarian	5	one-to-one
DN	5	$\sigma_{bbox} = 0.4, \sigma_{label} = 0.5, N_{dn} = 50$
CDN	5	$\sigma_{bbox} = 0.4, \sigma_{label} = 0.5, N_{dn} = 50$
Hybrid Matching	5	$N = 100, T = 300, K = 6$
Distinct Queries	5	$N = 100, \beta_{IoU} = 0.8$
DQM (ours)	5	$N = 100, \lambda = 0.2$
	VerSe	
Hungarian	24	one-to-one
DN	24	$\sigma_{bbox} = 0.2, \sigma_{label} = 0.25, N_{dn} = 50$
CDN	24	$\sigma_{bbox} = 0.2, \sigma_{label} = 0.25, N_{dn} = 50$
Hybrid Matching	24	$N = 100, T = 300, K = 6$
Distinct Queries	24	$N = 200, \beta_{IoU} = 0.8$
DQM (ours)	24	$N = 200, \lambda = 0.2$
	WORD	
Hungarian	10	one-to-one
DN	10	$\sigma_{bbox} = 0.4, \sigma_{label} = 0.5, N_{dn} = 50$
CDN	10	$\sigma_{bbox} = 0.4, \sigma_{label} = 0.5, N_{dn} = 50$
Hybrid Matching	10	$N = 100, T = 300, K = 6$
Distinct Queries	10	$N = 200, \beta_{IoU} = 0.8$
DQM (ours)	10	$N = 200, \lambda = 0.2$
	Total-Segmentator	
Hungarian	19	one-to-one
DN	19	$\sigma_{bbox} = 0.4, \sigma_{label} = 0.5, N_{dn} = 50$
CDN	19	$\sigma_{bbox} = 0.4, \sigma_{label} = 0.5, N_{dn} = 50$
Hybrid Matching	19	$N = 100, T = 300, K = 6$
Distinct Queries	19	$N = 250, \beta_{IoU} = 0.8$
DQM (ours)	19	$N = 600, \lambda = 0.2$
	AMOS	
Hungarian	15	one-to-one
DN	15	$\sigma_{bbox} = 0.4, \sigma_{label} = 0.5, N_{dn} = 50$
CDN	15	$\sigma_{bbox} = 0.4, \sigma_{label} = 0.5, N_{dn} = 50$
Hybrid Matching	15	$N = 100, T = 300, K = 6$
Distinct Queries	15	$N = 300, \beta_{IoU} = 0.8$
DQM (ours)	15	$N = 400, \lambda = 0.2$

2.2 Methods setting

Table 3: Parameter setting of different backbones

ResNet3D	**FPN**	**SwinFPN**	**SwinUNETR**
start-channels: 32	start-channels: 24	start-channels: 24	feature-size: 48
input-conv-kernel-size: 7	kernel-size: 3	hidden-dim: 384	hidden-size: 768
input-conv-stride: 2	stride: 2	mlp-dim: 1024	mlp-dim: 3072
input-conv-padding: 3	padding: 1	num-heads: [3, 6, 12, 24]	num-heads: 12
conv-kernel-size: 3	bias: False	pos-encoding: sine	pos-embed: 'conv'
conv-stride: 2	norm-name: 'instance'	attn-drop-rate: 0.0	proj-type: 'conv'
		drop-path-rate: 0.1	norm-name: 'instance'
ResNet3D-50		drop-rate: 0.0	dropout-rate: 0.0
num-layers=[3, 4, 6, 3]		qkv-bias: False	qkv-bias: False
		strides: [[1, 1, 1], [2, 2, 2],	conv-block: True
ResNet3D-101		[2, 2, 2], [2, 2, 2],	res-block: True
num-layers=[3, 4, 23, 3]		[2, 2, 2], [2, 2, 2]]	

References

[1] J. Ma, Y. Zhang, S. Gu, C. Zhu, C. Ge, Y. Zhang, X. An, C. Wang, Q. Wang, X. Liu *et al.*, "Abdomenct-1k: Is abdominal organ segmentation a solved problem?" *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, pp. 6695–6714, 2021.

[2] Y. Ji, H. Bai, C. Ge, J. Yang, Y. Zhu, R. Zhang, Z. Li, L. Zhanng, W. Ma, X. Wan *et al.*, "Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation," *Advances in Neural Information Processing Systems*, vol. 35, pp. 36 722–36 732, 2022.

[3] B. Wittmann, F. Navarro, S. Shit, and B. Menze, "Focused decoding enables 3d anatomical detection by transformers," *Machine Learning for Biomedical Imaging*, vol. 2, pp. 72–95, 2023.

[4] J. Wasserthal, M. Meyer, H.-C. Breit, J. Cyriac, S. Yang, and M. Segeroth, "Totalsegmentator: robust segmentation of 104 anatomical structures in ct images," *arXiv preprint arXiv:2208.05868*, 2022.

[5] A. Sekuboyina, M. E. Husseini, A. Bayat, M. Löffler, H. Liebl, H. Li, G. Tetteh, J. Kukačka, C. Payer, D. Štern *et al.*, "Verse: a vertebrae labelling and segmentation benchmark for multi-detector ct images," *Medical image analysis*, vol. 73, p. 102166, 2021.

[6] M. T. Löffler, A. Sekuboyina, and et al., "A vertebral segmentation dataset with fracture grading," *Radiology: Artificial Intelligence*, vol. 2, no. 4, p. e190138, 2020.

[7] H. Liebl, D. Schinz, A. Sekuboyina, L. Malagutti, M. T. Löffler, A. Bayat, M. El Husseini, G. Tetteh, K. Grau, E. Niederreiter *et al.*, "A computed tomography vertebral segmentation dataset with anatomical variations and multi-vendor scanner data," *Scientific data*, vol. 8, no. 1, p. 284, 2021.

Table 4: Parameter setting of different necks

RetinNet	**Foc. Dec.**	**D-DETR**	**D-DETR**
cls-channels: 384	num-feature-scales: 3	num-feature-scales: 3	num-feature-scales: 3
seg-channels: 24	pos-encoding: sine	pos-encoding: sine	pos-encoding: sine
head-channels: 192	hidden-dim: 384	hidden-dim: 384	hidden-dim: 384
learnable-scale: True	dropout: 0.1	dropout: 0.1	dropout: 0.1
num-candidates: 4	num-heads: 8	num-heads: 6	num-heads: 6
positive-fraction: 0.33	mlp-dim: 1024	mlp-dim: 1024	mlp-dim: 1024
pool-size: 20	num-layers: 4	num-layers: 4	num-layers: 4
min-neg: 1	restrict-attn: True	num-points: 4	
	obj-self-attn: False		
Post-processing			
score-thresh: 0	*Anchor Params.*		
topk-candidates: 10000	gen-dynamic-offset: True		
remove-small-boxes: 0.01	gen-offset: 0.1		
	offset-pred: True		
Anchor Params.	max-anchor-pred-offset: 0.1		
stride: 1			
anchors-per-position: 27			
depth: [[6, 10, 32], [12, 20, 64], [24, 40, 128] [48, 80, 256]]			
height: [[8, 11, 16], [16, 22, 32], [32, 44, 64], [64, 88, 128]]			
width: [[6, 8, 11], [12, 16, 22], [24, 32, 44], [48, 64, 88]]			

[8] X. Luo, W. Liao, J. Xiao, J. Chen, T. Song, X. Zhang, K. Li, D. N. Metaxas, G. Wang, and S. Zhang, "Word: A large scale dataset, benchmark and clinical applicable study for abdominal organ segmentation from ct image," *arXiv preprint arXiv:2111.02403*, 2021.

Table 5: mAP_{nndet} scores of the organ detection methods across five 3D CT datasets.

Method	AbdomenCT-1K	WORD	TS	AMOS	VerSe
nnU-Net-Plus	88.2	73.2	85.4	78.8	84.2
Retina U-Net	96.5	78.4	88.7	61.5	89.5
FocusedDec	98.7	90.4	91.9	87.8	90.6
Transoar	99.3	95.9	92.1	87.7	93.0
Organ-DETR	99.6	97.1	94.4	93.4	97.3

Table 6: AP_{50} scores of the organ detection methods across five 3D CT datasets.

Method	AbdomenCT-1K	WORD	TS	AMOS	VerSe
nnU-Net-Plus	86.7	64.9	75.2	60.3	71.8
Retina U-Net	95.4	71.6	74.6	58.4	75.3
FocusedDec	94.3	72.3	82.7	65.7	81.2
Transoar	96.9	85.7	80.5	68.9	83.1
Organ-DETR	98.3	90.5	85.6	80.6	90.7