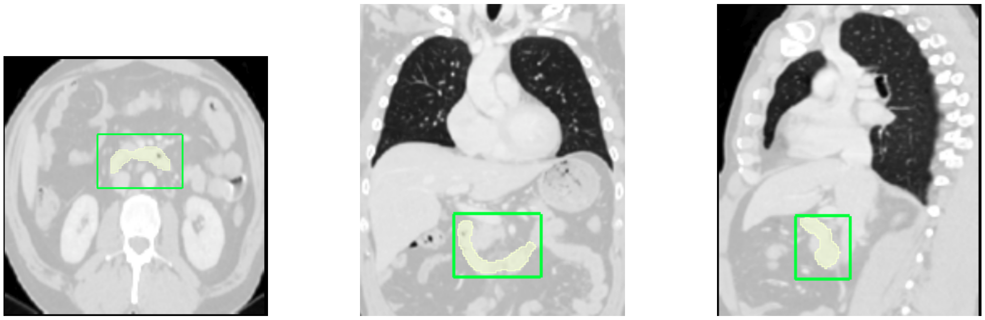
B SUPPLEMENTARY MATERIAL: ORGAN-DETR: 3D ORGAN DETECTION TRANSFOMER WITH MULTISCALE ATTENTION AND DENSE QUERY MATCHING

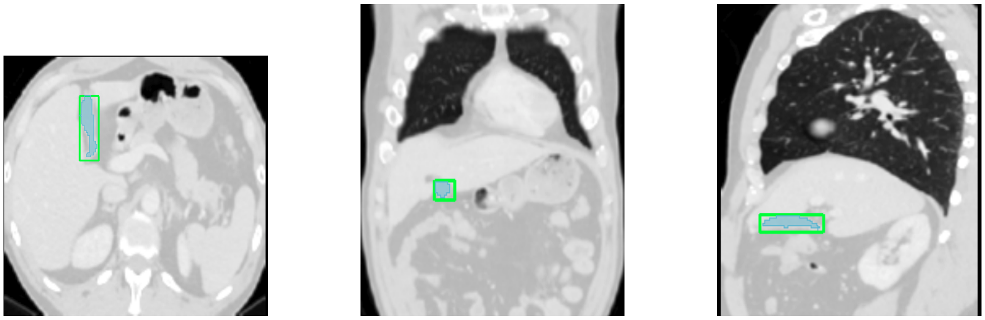
CONTENTS

B.1 CHALLENGES FOR ORGAN DETECTION IN CT

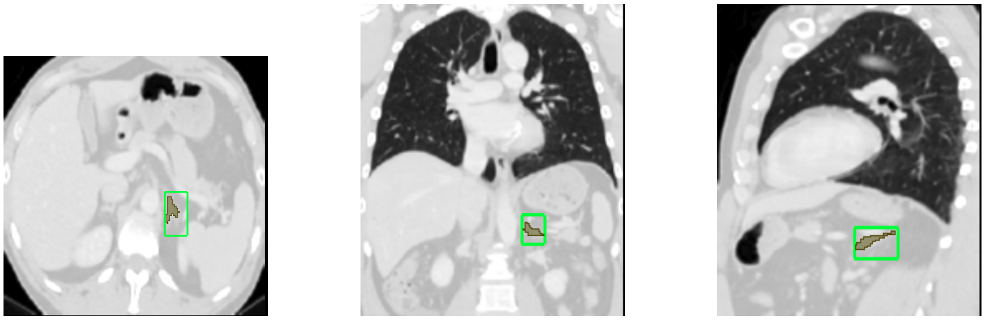
We visualize challenges for organ detection in a CT sample in Figure 7. The current challenges of CT data and detection can be categorized into four groups, as detailed below. *i)* Similar voxel intensity values in organs and neighboring tissues make the feature representation difficult to distinguish those based solely on voxel intensities from the same scale level. *ii)* Accurately outlining the border of individual organ structures is another challenge in CT data due to the unclear boundaries between adjacent organs and soft tissues. *iii)* Furthermore, the close proximity of organs within the human body and the potential for overlap in medical images pose challenges for detection in distinguishing between organs, particularly when they exhibit similar intensity levels. *iv)* Last but not least, individual anatomy may exhibit substantial variations, including the differences in the size of organs, shape, and positioning of those regarding each other. This adds another complexity to organ detection methods.



(a) Organ: Duodenum



(b) Organ: Gallbladder



(c) Organ: (Left) Adrenal Gland

Figure 7: Representation of three different organs in a 3D CT scan of Total-Segmentator (scan 11) from different viewpoints (left to right: axial, coronal, and sagittal). Precisely identifying organs presents a significant challenge due to 1) **proximity and overlap**: Organs in the human body are very close to each other and may overlap in medical images, which makes it difficult for automated systems to distinguish one organ from another, especially when they share similar intensities; 2) **fuzzy boundaries**: The boundaries between neighboring organs and soft tissues are not always well-defined, making it difficult to precisely delineate the borders of individual structures, 3) **similar intensity voxels**: Organs often have similar pixel or voxel intensities as the surrounding tissues. This similarity in intensity makes it challenging to differentiate organs from their neighboring structures solely based on pixel values in medical images; and 4) **inter-patient variability**: Each individual's anatomy can vary significantly. The size, shape, and even the position of organs can differ from one person to another. This adds another complexity layer when developing automated organ detection methods.

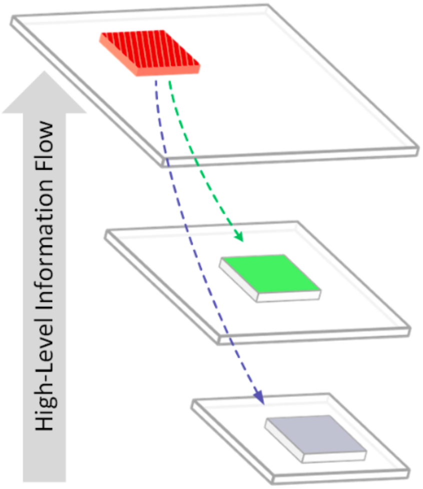


Figure 8: MSA is equipped with dual attention modules: a self-attention mechanism to capture short-range information within a layer and a cross-attention mechanism to capture long-range information between layers (shown by the green and blue arrows).

Table 3: Specification of five CT datasets publicly accessible and utilized in this study

Dataset	Size	#Train / #Val. / #Test	#Organs
AbdomenCT-1K (Ma et al., 2022)	224×224×96	732 / 126 / 116	5
AMOS (Ji et al., 2022)	256×256×128	166 / 22 / 52	15
WORD (Xiangde Luo & Zhang, 2022)	224×224×160	98 / 15 / 29	10
Total-Segmentator (Wasserthal et al., 2022)	160×160×256	113 / 21 / 29	19
VerSe (Sekuboyina et al., 2020)	64×64×256	261 / 37 / 74	24

B.2 INFORMATION FLOW IN MULTISCALE ATTENTION

Figure 8 visualizes the concept of the introduced Multiscale Attention (MSA). DETR-like methods often opt for high-level features due to the computational complexity, resulting in reduced spatial resolution. This constraint hinders the effective utilization of the Transformer-based features within the scope of detection tasks. In response to these challenges, MSA strategically captures a wide spectrum of long- and short-range feature patterns within and between layers with a dual attention mechanism.

Table 4: 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$\}$

Table 5: Summary of shape and voxel spacing characteristics in the preprocessed CT datasets

Dataset	Data Resolution	Voxel spacing
AbdomenCT-1K	(224,224,96)	(1.14,0.75,1.13) / (1.67,1.22,2.78) / (1.37,0.90,2.19)
WORD	(224,224,160)	(1.02,0.61,2.25) / (1.69,1.51,3.28) / (1.21,0.81,2.70)
Total-Segmentator	(160,160,256)	(1.31,0.67,0.16) / (2.59,2.09,2.56) / (1.88,1.40,2.22)
AMOS	(256,256,128)	(0.83,0.50,1.31) / (1.22,0.94,3.79) / (1.02,0.66,3.13)
VerSe	(64,64,256)	(3,3,3) / (3,3,3) / (3,3,3)

B.3 DATASETS AND DATA PREPARATION

The proposed organ detector is assessed using five publicly available CT datasets, see Table 3. All these datasets provide segmentation labels. During the training process, axis-aligned bounding boxes are extracted from the segmentation maps for each class. These bounding boxes serve as ground truth bounding boxes for training the object detector. An overview of statistics related to CT scans' size, voxel spacing, and intensity percentiles is reported in Table 5.

ABDOMENCT-1K

AbdomenCT-1K (Ma et al., 2022) comprises a total of 1112 abdominal CT scans gathered from various medical centers. In this dataset, each scan has voxel-wise segmentation labels for four organs. These CT scans share a common axial pixel resolution of 512×512. The slice thickness varies between $1.25mm$ and $5.0mm$. For this particular project, a preprocessed version of the dataset was utilized. In the preprocessing step, the data was resampled to have an anisotropic voxel spacing of 2mm along each axis. Furthermore, scans containing missing organs, such as the kidney, were excluded from the dataset, resulting in 975 CT scans. Additionally, a subset consisting of 160 samples was created for development purposes. The original segmentation labels in this dataset covered the liver, kidney, spleen, and pancreas. To enhance the dataset's utility, an extra label was introduced to differentiate between the left and right kidney (as indicated in Table 4). To achieve this, a script was developed to determine the centers of both kidneys, separate them using a sagittal plane, and then relabel the left kidney accordingly. To ensure uniformity across all scenes, the dataset's preprocessing involved registering all CT scans to the first scan in the dataset. This registration process ensured that the body's orientation in all CT scans was consistent. Metadata contained in the NIfTI files further confirmed the correctness of the orientation.

WORD

The WORD (Whole abdominal Organs Dataset) dataset (Xiangde Luo & Zhang, 2022) comprises 150 CT instances, all obtained from the same medical center and imaging device. This dataset provides segmentation labels for 16 anatomical structures. Each CT sample contains a variable number of slices, ranging from 159 to 330, with a consistent resolution of 512×512 pixels. The axial in-plane spacing is $0.976mm \times 0.976mm$, and the spacing between slices varies from $2.5mm$ to $3.0mm$. For the purpose of organ detection, the dataset excluded the femur heads. Out of the 14 remaining anatomical structures, 11 were retained, while the gallbladder, adrenal glands, and esophagus were excluded for simplicity. Additionally, the labels for the rectum and small bowel were combined into a single label, resulting in a total of 10 organs as outlined in Table 4.

TOTAL-SEGMENTATOR

The Total-Segmentator dataset (Wasserthal et al., 2022) consists of 1204 CT instances collected from various medical centers, encompassing various field-of-views (FOVs). This dataset provides segmentation labels for 104 anatomical structures, including bones, muscles, organs, and vessels. Each CT sample varies in composition, containing anywhere from 77 to 486 slices, with variable in-plane axial pixel resolution. Additionally, each volume in this dataset has a consistent isotropic spacing of $1.5mm$. A metadata table is available, containing information about FOV categories. A subset named 'Thorax-Abdomen-Pelvis (TAP)' has been defined to evaluate the object detector, comprising 19 organs as specified in Table 4. In this subset, certain organs like the heart and lungs had separate labels for different parts, which were consolidated into labels such as heart, left lung, and right lung for evaluation purposes.

AMOS

The AMOS (Abdominal Multi-Organ Segmentation) dataset (Ji et al., 2022) consists of a total of 500 CT samples obtained from different medical centers and imaging devices. This dataset provides segmentation labels for 15 organs, as specified in Table 4. Each CT instance varies in composition, containing between 67 and 369 slices. Out of these instances, 385 have an axial in-plane resolution of 512×512 pixels, while 115 have a resolution of 768×768 pixels. The median voxel spacing for the images is $0.67mm \times 0.67mm \times 5.0mm$, with a minimum of $0.45mm \times 0.45mm \times 1.25mm$ and a maximum of $1.07mm \times 1.07mm \times 5.0mm$. It is worth noting that only the training and validation subsets of this dataset have been used to create training, validation, and test datasets because the labels for the original test dataset are not publicly available.

VERSE

The VerSe (Vertebrae Segmentation) dataset (Sekuboyina et al., 2020; Löffler et al., 2020; Liebl et al., 2021) comprises a total of 374 CT images. These visual scans originate from CT scanners produced by four different manufacturers. Additionally, the dataset encompasses a wide range of field-of-views (FOVs) and includes various abnormalities such as fractures, metallic implants, and foreign materials. Specifically, segmentation labels are available for 26 vertebrae; however, to evaluate the organ detector, only 24 Vertebrae are considered. Vertebraes L6 and T13 were excluded due to their limited representation in the dataset; they were entirely absent from the validation and test datasets. Each CT image consists of a variable number of slices, ranging from 34 to 2023, with resolutions falling within the range of [103, 144] to [960, 2048]. The in-plane spacing varies from $0.195mm \times 0.195mm$ to $1.675mm \times 1.675mm$, and the slice thickness varies from $0.4mm$ to $5.0mm$. Although the VerSe dataset does not contain any organ-related data, it is still valuable because it presents a challenging organ detection task for CT instances.

B.4 PREPROCESSING AND AUGMENTATION

AbdomenCT-1K: Preprocessing 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 all datasets, ensuring uniformity and compatibility in the dataset preparation process.

WORD: 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.

Total-Segmentator: 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 4) since the original dataset defines 104 anatomical structures.

AMOS: Only the liver, stomach, spleen, urinary bladder, and prostate/uterus were considered during the organ cropping process, while the aorta and esophagus were excluded. Subsequently, 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 (Wittmann et al., 2023) for tests. The same set of organs and preprocessing steps were used for this dataset to keep comparability.

VerSe: The VerSe dataset contains varying FOVs, so the preprocessing of its CT scans differs from the preprocessing of the other datasets. 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 preprocessed VerSe dataset comprises a total of 372 CT samples.

The preprocessing of scans involved several steps to prepare them for further analysis:

- Normalization: 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.
- Clipping Intensities: Any intensities outside the normalized range were clipped to 0 or 1.
- Data Augmentations: To enhance the generalizability of the models, several augmentations were applied with a probability of 50%. These augmentations included:
  - Intensity Scaling and Shifting: The intensity values were scaled and shifted, each with a maximum variation of up to 10%.
  - Rotation: The scans were rotated by angles ranging from -5 degrees to +5 degrees.
  - Random Translations: Scans were randomly translated, with a maximum displacement of up to 10%.
  - Random Zoom: A random zoom, ranging from -10% to +10%, was applied with a probability of 50%.

These preprocessing and augmentation techniques were employed to make the dataset more robust and diverse, thereby improving the generalization capabilities of the models used for analysis.

B.5 EXPERIMENTAL SETUP

General setting: All the methods benefited from the AdamW algorithm[2]. For the neck (i.e., Detection Transformer), a learning rate of 2e-4 was applied, while for the backbone, it was set at 2e-5, with a weight decay of 1e-4. To manage the learning rate schedule, the StepLR scheduler[3] was employed, with a step size of 1250, throughout a span of 2000 epochs. A batch size of 2 was utilized for experiments across all datasets. The loss weights for classification, bounding box's IoU, bounding box's GIoU, and segmentation were configured as 2, 5, 2, and 2, respectively. Among the competing techniques, RetinaNet and Focused Decoder benefit from segmentation loss in the backbone. Specifically, they utilize the first layer with an identical resolution to the input CT data for segmentation purposes.

Backbones: In this study, the feature embedding size (f_e) remains constant at 384. The channel configuration for the ResNet model is defined as [32, 64, 256, 512, 1024], with layer settings configured as [3, 4, 6, 3] for ResNet-50 and [3, 4, 23, 3] for ResNet-101. The FPN backbone comprises 5 layers, initiating with a channel size of 24 and doubling it with each subsequent layer. The Swin Transformer was incorporated into the backbones' encoder primarily because of its promising performance. In this configuration, a window size of 5 was used. The second and subsequent layers had a stride of 2, whereas the first layer had a stride of 1. The depth was consistently set at 2, and the number of heads varied across layers, with values of [3, 6, 12, 24]. Additionally, a dropout rate of 0.2 was applied.

Detection Transformer: Organ-DETR, SwinFPN, and Transoar incorporate the decoder of Deformable DETR in their architecture, with the following configuration: 3 highest feature maps, embedding dimension of 384, 0.1 dropout rate, 6 attention heads, and 3 decoder layers. The focused Transformer unit in Focused Decoder maintained the same configuration as previously described. Additionally, the anchor offset was established at 0.1.

---

[2] `https://pytorch.org/docs/stable/generated/torch.optim.AdamW.html`

[3] `https://pytorch.org/docs/stable/generated/torch.optim.lr_scheduler.StepLR.html`

Matchers' setup: Tables 14 through 18 provide detailed parameter settings for each matching approach across all datasets. It is worth noting that the cost weights for classification, bounding box IoU, and bounding box GIoU were established at values of 2, 5, and 2, respectively.

EVALUATION METRICS

According to the COCO reference[4], 'mAP COCO', denoted by mAR or AP, computes average precision values at ten different IoU thresholds, ranging from 0.5 to 0.95 with increments of 0.05, i.e., $\mathbb{T} = \{0.5, 0.55, \ldots, 0.95\}$:

$$mAP = \frac{1}{|\mathbb{T}|} \sum_{t=<\mathbb{T}>} AP_t. \tag{18}$$

mAP is averaged over all categories. Likewise, mAR COCO computes AR (Average Recall) values at different IoU with thresholds in $\mathbb{T} = \{0.5, 0.55, \ldots, 0.95\}$:

$$mAR = \frac{1}{|\mathbb{T}|} \sum_{t=<\mathbb{T}>} AR_t. \tag{19}$$

AP_{75} is a strict metric that computes AP for predicted bounding boxes with the specific IoU threshold of 0.75. While incorporating additional metrics like localization error in millimeters could provide further assessment, our study emphasizes organ detection as a general task, demonstrating its applicability across various CT datasets with distinct characteristics. Hence, we believe the aforementioned metrics are well-suited for evaluating detection tasks.

B.6 EXPERIMENTAL RESULTS

B.6.1 ABLATION STUDY ON MSA PARAMETERS

MSA introduces a set of hyperparameters encompassing the size of voxel patches, the number of heads, drop rate, and depth. The influence of each of these parameters is detailed in Table 6, with an analysis conducted on the WORD dataset. Additionally, we have conducted an in-depth investigation into the impact of MSA's scale levels and have provided averaged results across all five datasets, reported in the table. Considering all metrics, it is advisable to opt for voxel patches of size 4 or 2 when using MSA. AP_{75} score associated with voxel patches of size 6 shows that larger voxel patches are unfavorable. The significance of MSA has been demonstrated in the preceding sections. Here, we underline the link between the number of scales employed in MSA and the resulting scores. Table 6 reveals that incorporating additional scale levels with cross-attention mechanisms significantly enhances the performance of Organ-DETR. Notably, using two cross-attentions results in a remarkable performance gain of +6.0 mAP, underscoring the efficacy of involving a broader range of voxel patches from various scales to enhance the detection method's overall performance. The table also suggests 64 heads and a depth of 4.

Table 6: Ablation study on different parameters of MSA on WORD and other datasets.

Parameter	mAP	mAR	AP_{75}
Voxel Patch Size			
$2 \times 2 \times 2$	51.1	57.9	55.0
$4 \times 4 \times 4$	51.6	58.3	55.0
$6 \times 6 \times 6$	51.5	57.9	53.0
Number of Scale Levels (avg. across all 5 datasets)			
1 (w/o cross-att.)	44.7	39.6	51.6
2 (1 cross-att. level)	48.3	45.5	55.6
3 (2 cross-att. levels)	54.3	61.4	56.0
Number of Heads			
16	50.1	57.9	50.8
32	51.6	58.3	55.0
64	53.3	59.1	57.2
Depth			
2	51.6	58.3	55.0
3	51.3	58.2	54.8
4	51.9	59.2	54.0
Drop rate			
0	51.1	57.9	55.0
0.1	50.9	56.5	59.4
0.3	52.0	58.3	53.5

---

[4] `https://cocodataset.org/#detection-eval`

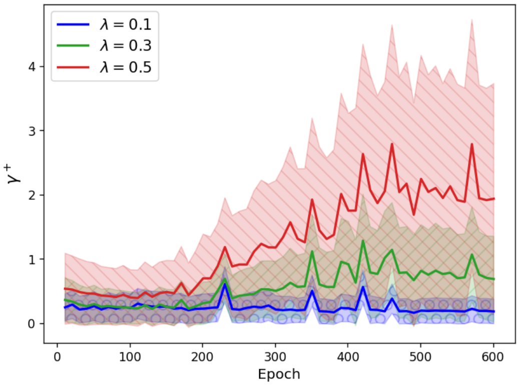


Figure 9: Standard deviation and mean (solid line) of gradient descent results for positive queries in a sample from the WORD dataset

B.6.2 RESULTS ON GRADIENT NORM

Figure 9 depicts the gradient norm values of positive queries within the DQM method, showcasing varying values of the parameter λ. The graph reveals a clear trend: higher values of λ correspond to increased gradient values, thereby expediting the training process. Higher values of λ may also lead to a higher incidence of false positive queries. However, the multiscale segmentation loss framework effectively controls and mitigates this potential issue. It is worth noting that equation 6 is calculated based on the assumption of uniform gradient values for all positive queries, a condition that may not hold in practical scenarios. Consequently, some variance in the results depicted in the figure compared to those obtained in theory is expected, but the overall trend substantiates the conclusion from the theoretical analysis.

B.6.3 COMPUTATIONAL COST

All experiments on the WORD, Total-Segmentator, and VerSe datasets were conducted on an NVIDIA A100 with 40GB VRAM. Due to their large data size, we employed an NVIDIA A100 with 80GB VRAM for running the experiments on the AbdomenCT-1K and AMOS. In the analysis of computational cost, we assessed the cost-effectiveness of the detection methods using four key metrics: the total number of parameters, the number of floating-point operations (FLOPs), the training duration in hours on a GPU, and the inference frames per second (FPS). The results for WORD and VerSe are reported in Tables 7 and 8, respectively. While training Organ-DETR takes longer, it exhibits superior inference speed compared to alternative techniques. Note that we did not report the cost of Focused Dec on the VerSe dataset since it requires fixed FoV, so VerSe is not applicable. Additionally, it is worth highlighting that Organ-DETR possesses the highest number of parameters among the considered models. This observation underscores the remarkable computational efficiency achieved by the Organ-DETR framework.

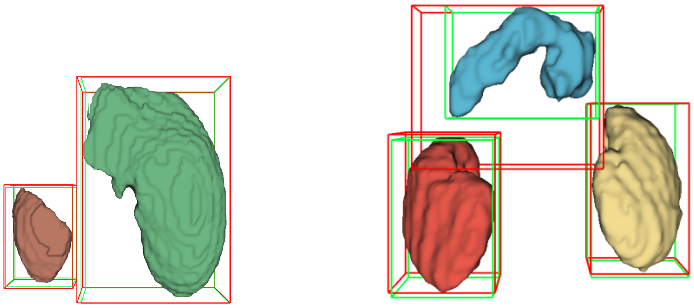
Table 7: Comparative results of organ detection techniques in terms of computational cost on WORD dataset

Method	Backbone	Transformer	Params (#M)	FLOPs (#G)	Training (GPU hours)	Inference FPS
RetinaNet	FPN	Retina U-Net	51	1583	17	12
-	FPN	DETR	44	612	21	5
Focused Dec.	FPN	Foc. Dec.	43	373	20	14
SwinFPN	FPN + Swin	D-DETR	73	638	22	9
Transoar	FPN	D-DETR	53	583	20	12
Organ-DETR	FPN	MSA + D-DETR	84	629	26	18

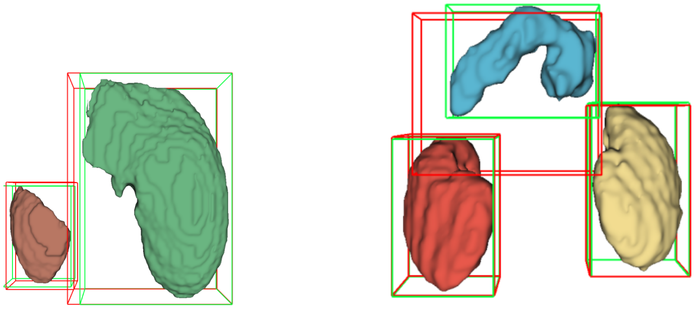
Table 8: Comparative results of organ detection techniques in terms of computational cost on the VerSe dataset

Method	Backbone	Transformer	Params (#M)	FLOPs (#G)	Training (GPU hours)	Inference FPS
RetinaNet	FPN	Retina U-Net	53	203	13	59
-	FPN	DETR	44	63	8	32
Focused Dec.	FPN	Foc. Dec.	–	–	–	–
SwinFPN	FPN + Swin	D-DETR	63	66	11	34
Transoar	FPN	D-DETR	53	56	9	42
Organ-DETR	FPN	MSA + D-DETR	84	108	39	41

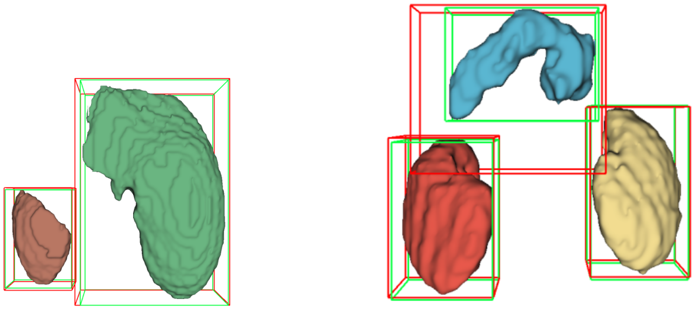
B.6.4 VISULAISION COMPARISON



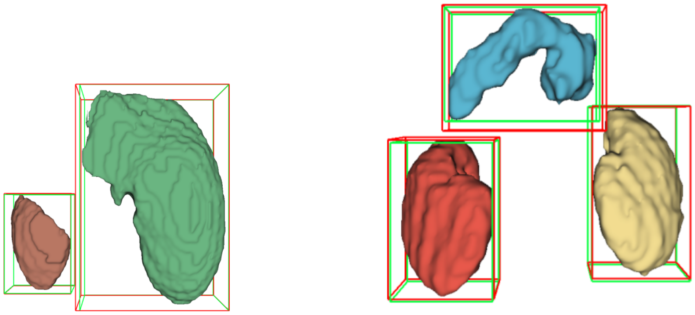
(a) RetinaNet
(Spleen: 0.89, Liver: 0.94, Right Kidney: 0.87, Pancreas: 0.53, Left Kidney: 0.94)



(b) Focused Decoder
(Spleen: 0.84, Liver: 0.88, Right Kidney: 0.94, Pancreas: 0.49, Left Kidney: 0.94)

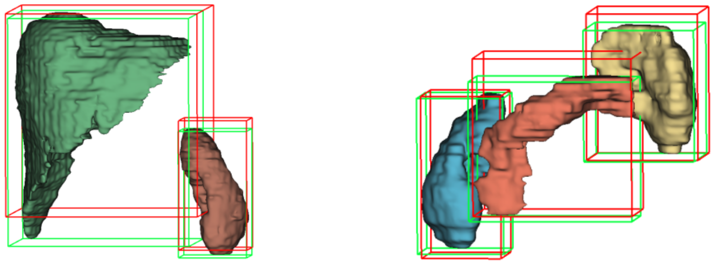


(c) Transoar (FPN)
(Spleen: 0.91, Liver: 0.94, Right Kidney: 0.89, Pancreas: 0.50, Left Kidney: 0.93)

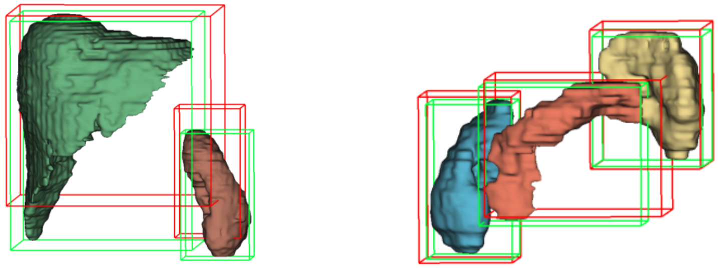


(d) Organ-DETR (FPN)
(Spleen: 0.96, Liver: 0.94, Right Kidney: 0.96, Pancreas: 0.83, Left Kidney: 0.90)

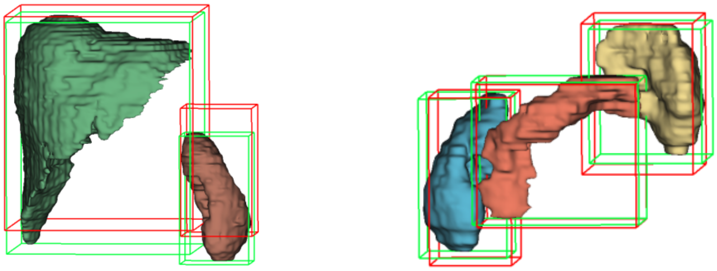
Figure 10: Visual comparison of bounding box predictions for various organ detection techniques in a subset of the AbdomenCT-1K dataset, along with reported IoU scores for each organ. Note that due to the limitations of 3D visualization, some fine details may be lost in 2D visualization.



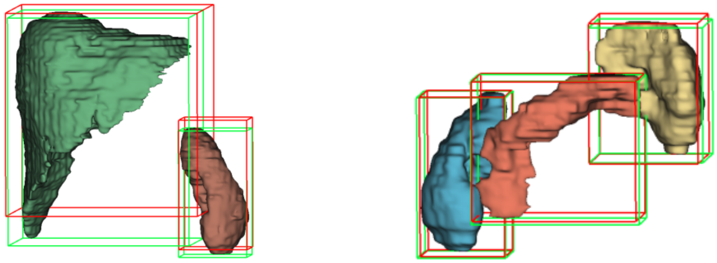
(a) RetinaNet
(Liver: 0.76, Left Kidney: 0.76, Right Kidney: 0.91, Pancreas: 0.56, Spleen: 0.77)



(b) Focused Decoder
(Liver: 0.76, Left Kidney: 0.49, Right Kidney: 0.60, Pancreas: 0.61, Spleen: 0.69)

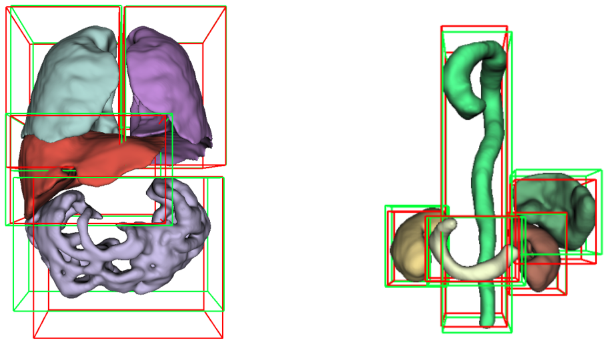


(c) Transoar (with FPN backbone)
(Liver: 0.83, Left Kidney: 0.62, Right Kidney: 0.81, Pancreas: 0.67, Spleen: 0.69)

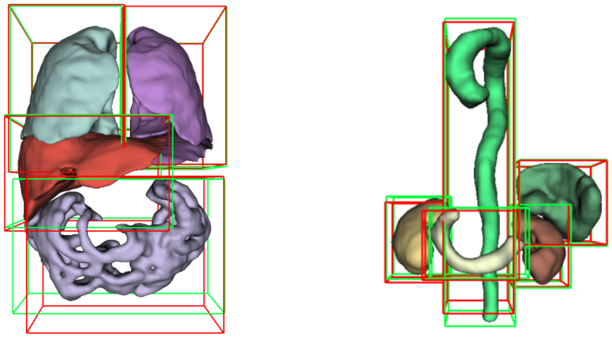


(d) Organ-DETR (with FPN backbone)
(Liver: 0.85, Left Kidney: 0.79, Right Kidney: 0.93, Pancreas: 0.68, Spleen: 0.77)

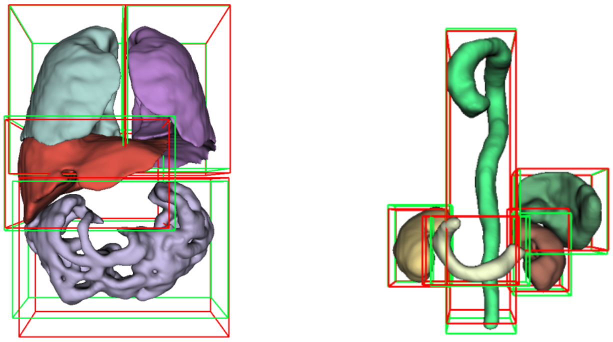
Figure 11: Visual comparison of bounding box predictions for various organ detection techniques in a subset of the WORD dataset, along with reported IoU scores for each organ. Note that due to the limitations of 3D visualization, some fine details may be lost in 2D visualization.



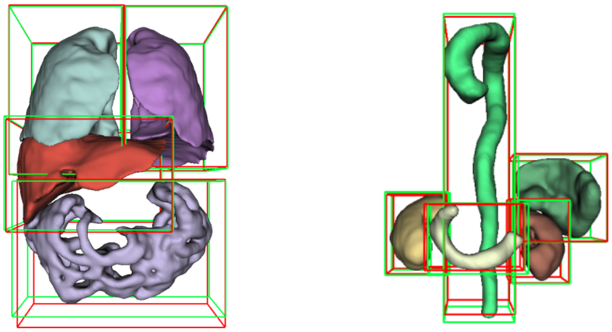
(a) RetinaNet
(Avg. IoUs of depicted organs-Left: 0.88, Right: 0.81)



(b) Focused Decoder
(Avg. IoUs of depicted organs-Left: 0.90, Right: 0.86)

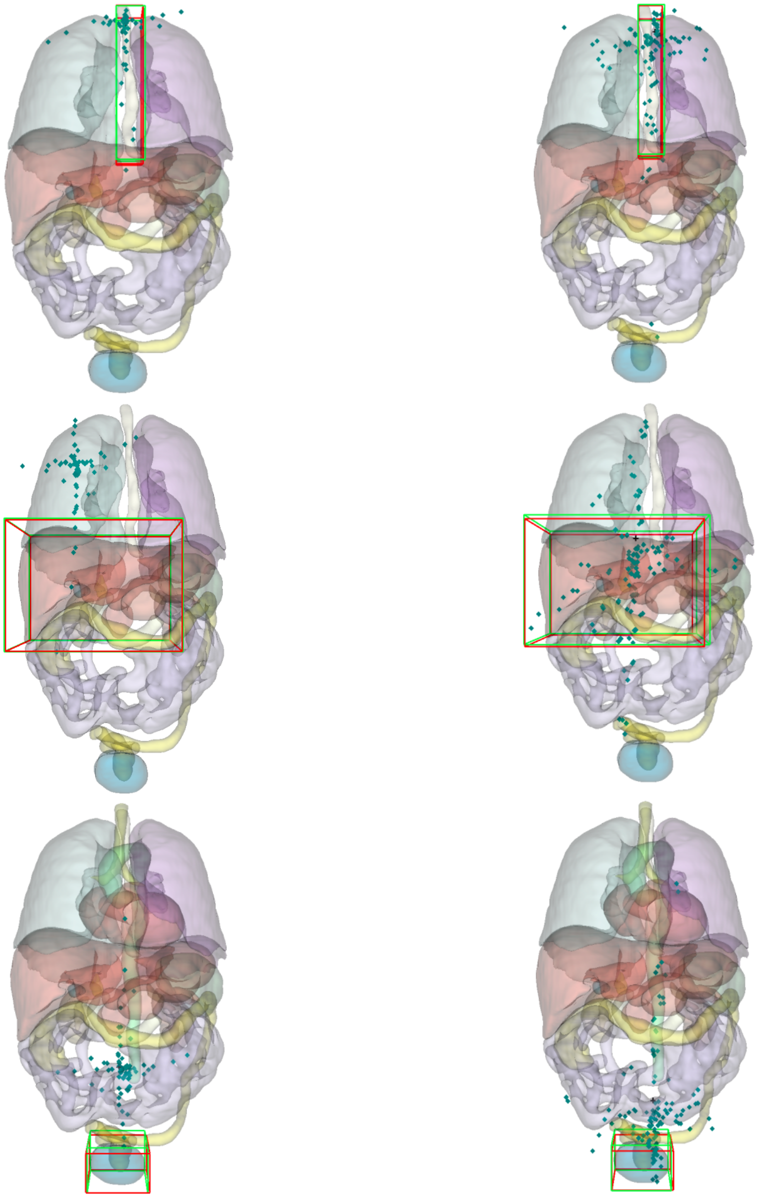


(c) Transoar (with FPN backbone)
(Avg. IoUs of depicted organs-Left: 0.89, Right: 0.81)



(d) Organ-DETR (with FPN backbone)
(Avg. IoUs of depicted organs-Left: 0.92, Right: 0.88)

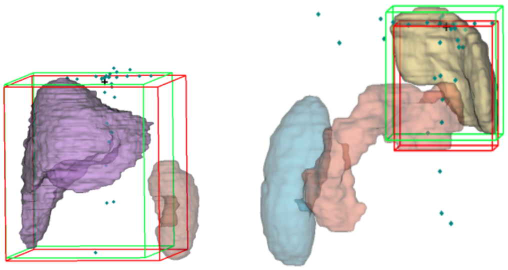
Figure 12: Visual comparison of bounding box predictions for various organ detection techniques in a subset of the Total-Segmentator dataset, along with reported IoU scores for each organ. Note that due to the limitations of 3D visualization, some fine details may be lost in 2D visualization.



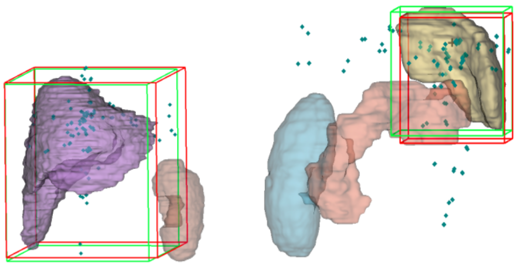
(a) Hungarian Matching
(Esophagus: 0.74, Liver: 0.87, Bladder: 0.15)

(b) DQM
(Esophagus: 0.81, Liver: 0.90, Bladder: 0.40)

Figure 13: Visualization of the sampling points' location of different organs in the last layer of the Transformer decoder for (a) Deformable DETR's Hungarian matching in Transoar; and (b) the proposed DQM of Organ-DETR on a Total-Segmentator's sample. Organs represented from top to bottom: Esophagus, liver, and bladder. Note that due to the limitations of 3D visualization, some fine details may be lost in 2D visualization.

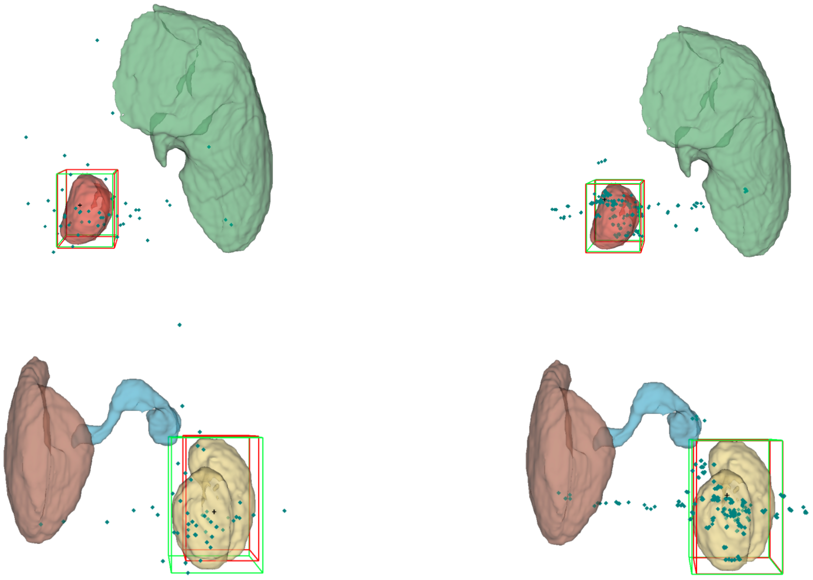


(a) Hungarian Matching
(Liver: 0.77, Left Kidney: 0.53)



(b) DQM
(Liver: 0.87, Left Kidney: 0.77)

Figure 14: Visualization of the sampling points' location of different organs in the last layer of the Transformer decoder for (a) Deformable DETR's Hungarian matching in Transoar; and (b) the proposed DQM of Organ-DETR on a WORD's sample. Note that due to the limitations of 3D visualization, some fine details may be lost in 2D visualization.



(a) Hungarian Matching
(Right Kidney: 0.53, Left Kidney: 0.89)

(b) DQM
(Right Kidney: 0.87, Left Kidney: 0.92)

Figure 15: Visualization of the sampling points' location of different organs in the last layer of the Transformer decoder for (a) Deformable DETR's Hungarian matching in Transoar; and (b) the proposed DQM of Organ-DETR on an AbdomenCT-1K's sample. Note that due to the limitations of 3D visualization, some fine details may be lost in 2D visualization.

B.6.5 MAIN RESULTS PER DATABASE

Table 9: Comparative results of organ detection techniques using consistent parameters on the AbdomenCT-1K dataset

Method	Backbone	Transformer	Matcher	mAP	mAR	AP_{75}	Query (#)	Params (#M)
RetinaNet	FPN	Retina U-Net	-	71.9	76.6	81.1	100	50.7
-	UNETR+Swin	D-DETR	Hung.	69.1	75.0	78.4	100	51.7
-	FPN	DETR	Hung.	64.9	70.4	72.1	100	43.7
Focused Dec.	FPN	Foc. Dec.	Hung.	60.3	67.4	66.8	135	36.1
SwinFPN	FPN + Swin	D-DETR	Hung.	76.3	81.9	85.9	100	38.2
Transoar	FPN	D-DETR	Hung.	76.4	81.5	85.0	100	53.4
Transoar	ResNet-50	D-DETR	Hung.	76.0	81.2	84.1	100	56.9
Organ-DETR	FPN	MSA + D-DETR	DQM	81.7	86.4	91.7	100	81.5
Organ-DETR	ResNet-50	MSA + D-DETR	DQM	80.0	84.9	84.9	100	53.5
Organ-DETR	ResNet-101	MSA + D-DETR	DQM	80.2	85.2	92.2	100	62.3

Table 10: Comparative results of organ detection techniques using consistent parameters on the WORD dataset

Method	Backbone	Transformer	Matcher	mAP	mAR	AP_{75}	Query (#)	Params (#M)
RetinaNet	FPN	Retina U-Net	-	34.5	39.4	29.2	100	51.4
-	UNETR+Swin	D-DETR	Hung.	39.0	47.2	30.1	100	51.6
-	FPN	DETR	Hung.	29.1	36.5	20.4	100	43.7
Focused Dec.	FPN	Foc. Dec.	Hung.	26.7	34.1	17.4	270	42.5
SwinFPN	FPN + Swin	D-DETR	Hung.	40.8	49.4	35.1	100	73.4
Transoar	FPN	D-DETR	Hung.	38.8	47.3	29.1	100	53.4
Transoar	ResNet-50	D-DETR	Hung.	34.4	42.6	24.0	100	56.9
Organ-DETR	FPN	MSA + D-DETR	DQM	52.0	58.3	53.5	100	84.3
Organ-DETR	ResNet-50	MSA + D-DETR	DQM	48.2	55.6	45.3	100	52.5
Organ-DETR	ResNet-101	MSA + D-DETR	DQM	48.4	55.1	44.2	100	62.3

Table 11: Comparative results of organ detection techniques using consistent parameters on the Total-Segmentator dataset

Method	Backbone	Transformer	Matcher	mAP	mAR	AP_{75}	Query (#)	Params (#M)
RetinaNet	FPN	Retina U-Net	-	35.8	41.7	27.5	100	52.7
-	UNETR+Swin	D-DETR	Hung.	37.5	43.0	31.7	100	51.7
-	FPN	DETR	Hung.	29.7	35.4	22.9	100	43.7
Focused Dec.	FPN	Foc. Dec.	Hung.	42.7	48.7	37.4	513	42.9
SwinFPN	FPN + Swin	D-DETR	Hung.	40.9	47.1	34.8	100	73.4
Transoar	FPN	D-DETR	Hung.	40.4	46.8	33.0	100	53.4
Transoar	ResNet-50	D-DETR	Hung.	37.6	43.9	31.9	100	56.9
Organ-DETR	FPN	MSA + D-DETR	DQM	49.5	55.4	49.6	190	84.4
Organ-DETR	ResNet-50	MSA + D-DETR	DQM	47.5	53.5	44.8	190	52.6
Organ-DETR	ResNet-101	MSA + D-DETR	DQM	47.0	53.1	44.4	190	62.3

Table 12: Comparative results of organ detection techniques using consistent parameters on the VerSe dataset

Method	Backbone	Transformer	Matcher	mAP	mAR	AP_{75}	Query (#)	Params (#M)
RetinaNet	FPN	Retina U-Net	-	46.3	55.4	51.0	100	53.4
-	UNETR+Swin	D-DETR	Hung.	36.3	45.2	25.8	100	51.9
-	FPN	DETR	Hung.	21.8	31.1	9.4	100	43.7
SwinFPN	FPN + Swin	D-DETR	Hung.	33.9	43.6	20.8	100	63
Transoar	FPN	D-DETR	Hung.	34.8	44.8	22.7	100	53.4
Transoar	ResNet-50	D-DETR	Hung.	35.8	44.9	24.4	100	56.9
Organ-DETR	FPN	MSA + D-DETR	DQM	55.1	61.8	62.7	240	84.4
Organ-DETR	ResNet-50	MSA + D-DETR	DQM	48.3	57.7	45.5	240	52.6
Organ-DETR	ResNet-101	MSA + D-DETR	DQM	49.5	56.6	46.8	240	62.4

Table 13: Comparative results of organ detection techniques using consistent parameters on the AMOS dataset

Method	Backbone	Transformer	Matcher	mAP	mAR	AP_{75}	Query (#)	Params (#M)
RetinaNet	FPN	Retina U-Net	-	27.9	31.6	24.6	100	52.2
-	UNETR+Swin	D-DETR	Hung.	20.1	27.8	9.1	100	51.7
-	FPN	DETR	Hung.	18.0	24.5	8.3	100	43.7
Focused Dec.	FPN	Foc. Dec.	Hung.	25.1	33.3	13.5	405	42.6
SwinFPN	FPN + Swin	D-DETR	Hung.	27.2	35.1	17.4	100	73.4
Transoar	FPN	D-DETR	Hung.	28.1	35.4	17.9	100	53.4
Transoar	ResNet-50	D-DETR	Hung.	24.1	85.5	31.2	100	56.9
Organ-DETR	FPN	MSA + D-DETR	DQM	36.2	43.6	29.2	150	84.3
Organ-DETR	ResNet-50	MSA + D-DETR	DQM	32.2	40.0	23.4	150	52.5
Organ-DETR	ResNet-101	MSA + D-DETR	DQM	32.9	40.7	24.3	150	62.3

B.6.6 MATCHING RESULTS PER DATABASE

Table 14: Comparative results of different matching techniques on the AbdomenCT-1K dataset

Method	#M	Hyperparameters	mAP	mAR	AP$_{75}$
Hungarian	5	one-to-one	76.4	81.4	93.0
DN	5	$\sigma_{bbox} = 0.4, \sigma_{label} = 0.5, N_{dn} = 50$	75.2	81.0	90.1
CDN	5	$\sigma_{bbox} = 0.4, \sigma_{label} = 0.5, N_{dn} = 50$	75.5	81.5	84.5
Hybrid Matching	5	$N = 100, T = 300, K = 6$	76.7	82.2	85.5
Matching with Distinct Queries	5	$N = 100, \beta_{IoU} = 0.8$	73.5	79.4	83.2
DQM (ours)	5	$N = 100, \lambda = 0.2$	78.9	83.9	93.0

Table 15: Comparative results of different matching techniques on the VerSe dataset

Method	#M	Hyperparameters	mAP	mAR	AP$_{75}$
Hungarian	24	one-to-one	35.5	45.1	23.6
DN	24	$\sigma_{bbox} = 0.2, \sigma_{label} = 0.25, N_{dn} = 50$	33.7	43.2	24.4
CDN	24	$\sigma_{bbox} = 0.2, \sigma_{label} = 0.25, N_{dn} = 50$	33.5	43.5	20.1
Hybrid Matching	24	$N = 100, T = 300, K = 6$	36.3	45.9	25.3
Matching with Distinct Queries	24	$N = 200, \beta_{IoU} = 0.8$	24.7	32.0	12.5
DQM (ours)	24	$N = 200, \lambda = 0.1$	36.4	45.0	23.6

Table 16: Comparative results of different matching techniques on the WORD dataset

Method	#M	Hyperparameters	mAP	mAR	AP$_{75}$
Hungarian	10	one-to-one	38.1	46.8	28.5
DN	10	$\sigma_{bbox} = 0.4, \sigma_{label} = 0.5, N_{dn} = 50$	37.5	46.0	28.3
CDN	10	$\sigma_{bbox} = 0.4, \sigma_{label} = 0.5, N_{dn} = 50$	38.8	47.4	27.6
Hybrid Matching	10	$N = 100, T = 300, K = 6$	41.7	49.5	34.8
Matching with Distinct Queries	10	$N = 200, \beta_{IoU} = 0.8$	37.7	45.8	26.8
DQM (ours)	10	$N = 200, \lambda = 0.2$	42.4	50.4	35.3

Table 17: Comparative results of different matching techniques on the Total-Segmentator dataset

Method	#M	Hyperparameters	mAP	mAR	AP$_{75}$
Hungarian	19	one-to-one	39.2	45.4	31.8
DN	19	$\sigma_{bbox} = 0.4, \sigma_{label} = 0.5, N_{dn} = 50$	38.5	45.1	32.2
CDN	19	$\sigma_{bbox} = 0.4, \sigma_{label} = 0.5, N_{dn} = 50$	38.7	45.4	33.1
Hybrid Matching	19	$N = 100, T = 300, K = 6$	37.3	44.3	31.1
Matching with Distinct Queries	19	$N = 250, \beta_{IoU} = 0.8$	28.8	34.5	21.0
DQM (ours)	19	$N = 600, \lambda = 0.2$	38.6	45.5	34.4

Table 18: Comparative results of different matching techniques on the AMOS dataset

Method	#M	Hyperparameters	mAP	mAR	AP$_{75}$
Hungarian	15	one-to-one	26.9	34.3	16.3
DN	15	$\sigma_{bbox} = 0.4, \sigma_{label} = 0.5, N_{dn} = 50$	26.1	34.4	14.9
CDN	15	$\sigma_{bbox} = 0.4, \sigma_{label} = 0.5, N_{dn} = 50$	26.4	34.1	17.6
Hybrid Matching	15	$N = 100, T = 300, K = 6$	27.0	34.5	15.9
Matching with Distinct Queries	15	$N = 300, \beta_{IoU} = 0.8$	19.5	26.8	7.4
DQM (ours)	15	$N = 400, \lambda = 0.2$	27.1	35.0	27.7