Q-Former Autoencoder: A Modern Framework for Medical Anomaly Detection Supplementary Material

Francesco Dalmonte[1,*] Emirhan Bayar[2,*] Emre Akbas[2,3] Mariana-Iuliana Georgescu[3]

[1]University of Bologna, Italy [2]Depart. of Computer Eng., METU, Ankara, Türkiye
[3]Helmholtz Munich, Germany

We provide additional implementation details in Section 1, additional experiments on LiverCT and RSNA in Section 2 and Section 3.

1. Implementation Details

This section provides an overview of the implementation details for our proposed framework, ensuring full reproducibility of our results. All experiments were conducted in PyTorch.

1.1. Hyperparameters

The main hyperparameters used for training and evaluation are detailed in Table 1 and Table 2, respectively.

Table 1. Training hyperparameters for the experiments.

Component	Parameter	Value
General	Seed	42, 7, 13, 65, 91 (mean of 5 runs are reported)
	Image Resolution (Resize)	224x224
	Batch Size	64
	Epochs	300
	Device	CUDA
Encoder	Pre-trained Model	ViT-Large (ViT-L/14) with register tokens
	Pre-training Method	DINOv2
	Frozen During Training	True
	Hidden States Used	Features from the 2nd and 4th to last blocks
	Final Projection In-Features	1024
	Final Projection Out-Features	768
Q-Former (Junction)	Number of Transformer Blocks	1
	Internal Dimension	768
	Output Dimension	768
	Number of Learnable Queries	784 (for 28x28 output patches)
	Attention Heads	8
	MLP Expansion Ratio	4.0
Decoder	Internal Dimension	768
	Depth (Number of Layers)	6
	Attention Heads	12
	Output Patch Size	8x8
	Number of Output Patches	28x28
	MLP Expansion Ratio	4.0
Optimization	Optimizer	Adam
	Learning Rate (Maximum)	8×10^{-5}
	Learning Rate Scheduler	OneCycleLR
Perceptual Loss	Pre-trained Perceptual Model	Masked Autoencoder (MAE) with ViT-Large Encoder
	Distance Metric	Cosine Distance
	Layers Used for Feature Extraction	From the 16th and 20th transformer blocks
	Multi-Scale Input Patch Sizes	32x32, 56x56

*Equal contribution

Table 2. Evaluation configuration for the experiments.

Component	Parameter	Value
General	Batch Size	64
	Test Data Augmentation	None (only resize and normalize)
Perceptual Metric	Pre-trained Perceptual Model	MAE with ViT-Large Encoder
	Distance Metric	Cosine Distance
	Layers Used for Feature Extraction	From the 12th, 16th, and 20th transformer blocks
	Multi-Scale Input Patch Sizes	16x16, 32x32, 56x56
Image-Level Score Aggregation	Spatial Aggregation per Feature Map	Max
	Cross-Feature Map Aggregation	Mean
Pixel-Level Map Aggregation	Cross-Feature Map Aggregation	Mean

1.2. Perceptual Loss Formulation

The training objective is to minimize a multi-scale perceptual loss. This loss is calculated in a three-step process:

Step 1: Feature Extraction. For an input image x and its reconstruction $\tilde{x}$, we extract feature maps from a set of pretrained perceptual models. We use multiple Masked Autoencoder (Masked AE) models, each distinguished by its input patch size $p \in P$. For each model, we select features from a set of transformer blocks $i \in I$. Let $\Phi_{i,p}(x)$ be the feature map of shape $C_i \times H_i \times W_i$ extracted from the i-th layer of the perceptual model with patch size p.

Step 2: Anomaly Map Calculation. For each selected feature map, we compute an intermediate anomaly map, $A_{i,p}$, by calculating the cosine distance between the features of the original image and its reconstruction at every spatial location (j, k).

$$A_{i,p}(j,k) = 1 - \frac{\Phi_{i,p}(x)_{j,k} \cdot \Phi_{i,p}(\tilde{x})_{j,k}}{\|\Phi_{i,p}(x)_{j,k}\|_2 \cdot \|\Phi_{i,p}(\tilde{x})_{j,k}\|_2}$$

This produces a set of single-channel anomaly maps, one for each combination of layer i and patch size p.

Step 3: Hierarchical Aggregation and Final Loss. The final loss is computed using a two-stage hierarchical aggregation. First, for each feature layer $i \in I$, we create a robust, layer-specific anomaly map, $A_{\text{combined},i}$, by performing an element-wise multiplication of its corresponding anomaly maps from all different patch-size models $p \in P$. This step enforces a strict consensus across multiple scales for each

feature level.

$$A_{\text{combined},i} = \prod_{p \in P} \text{Resize}_{(H,W)}(A_{i,p})$$

Second, the total loss $\mathcal{L}$ is calculated by averaging the mean value of each of these robust, layer-specific maps. This treats the error signal from each feature layer as an independent contribution to the total loss.

$$\mathcal{L}(x, \tilde{x}) = \frac{1}{|I|} \sum_{i \in I} \text{mean}\left(A_{\text{combined},i}\right)$$

For training, we use patch sizes $P = \{32, 56\}$ and features from the 16th and 20th transformer blocks of the Masked AE ViT-Large encoder.

1.3. Anomaly Score and Map Generation

During evaluation, we generate both an image-level scalar score for AUROC computation and a pixel-level anomaly map. Both start from the same set of intermediate anomaly maps, $A_{i,p}$, though computed using the evaluation configuration (Table 2). Let this evaluation set of maps be denoted by $\mathcal{A} = \{A_1, A_2, ..., A_N\}$.

Image-Level Anomaly Score Aggregation. To derive a single scalar score for each image, we perform a two-step aggregation:

Step 1: Spatial Aggregation. For each anomaly map $A_n \in \mathcal{A}$, we find the maximum pixel value. This value, s_n, represents the most severe reconstruction error detected by that specific feature map.

$$s_n = \max_{j,k}(A_n(j, k))$$

Step 2: Cross-Feature Aggregation. The final image-level score, A_{score}, is the mean of these maximum values, averaged over all N feature maps.

$$A_{\text{score}} = \frac{1}{N} \sum_{n=1}^{N} s_n$$

This method gives a robust score that is sensitive to strong local anomalies while benefiting from the diversity of features from different layers.

Pixel-Level Anomaly Map Generation. To generate a final 2D anomaly map, we use a different aggregation strategy that preserves spatial information. At each spatial location (j, k), we take the mean value across all N resized anomaly maps.

$$A_{\text{pixel-max}}(j, k) = \max_{n \in \{1..N\}}(A_n(j, k))$$

1.4. Training and Data Augmentation

The model is trained using the Adam optimizer with a `OneCycleLR` learning rate scheduler. To encourage the model to learn robust and generalizable representations of normal data, the following data augmentations are applied to the training set:

- **Random Resized Crop:** Images are cropped to a random size (90% to 100% of the original) and aspect ratio (80% to 120% of the original) before being resized to the final input dimension.
- **Random Rotation:** Images are rotated by a random angle between -10 and +10 degrees.
- **Random Vertical Flip:** Images are flipped vertically with a 50% probability.
- **Color Jitter:** The brightness and contrast of the images are randomly adjusted by a factor of up to 0.1.
- **Normalization:** Image pixel values are normalized to have a mean of 0.449 and a standard deviation of 0.226.

1.5. Datasets

We evaluate our method on four medical imaging datasets from the BMAD [4] benchmark, each presenting distinct anomaly detection challenges. Following standard protocol [4], we train separate models for each dataset rather than combining them, as the anomaly types differ fundamentally across domains (brain tumors, liver lesions, retinal fluids, and chest pathologies).

BraTS2021. The BraTS2021 [2, 3, 18] dataset contains brain MRI images from glioma patients. The dataset comprises 11,298 FLAIR modality images (7,500 training, 83 validation, 3,715 test) at 240×240 resolution, usually extracted from depth 60-100 of 3D volumes. The test set contains 3,075 anomalous and 640 normal slices. Anomalies consist of three tumor sub-regions: enhancing tumor (ET), tumor core (TC), and whole tumor (WT).

LiverCT. The LiverCT dataset combines BTCV [13] and LiTS [5] for liver anomaly detection. The training set comprises 1,542 normal liver slices from BTCV (anomaly-free abdominal CT scans), while test (1,493 slices with 660 anomalous and 833 normal) and validation (166 slices) sets are derived from LiTS containing both normal tissue and liver tumors. Images are 512×512 resolution with histogram equalization applied for enhancement.

RESC. The RESC dataset [11] contains 6,217 retinal OCT images (4,297 training, 115 validation, 1,805 test) at 512×1024 resolution for macular edema segmentation. The test set comprises 764 anomalous and 1,041 normal images. Anomalies include two lesion types: subretinal fluid (SRF) and pigment epithelium detachment (PED).

RSNA. The RSNA Pneumonia Detection dataset [27] contains 26,684 chest X-ray images at 1024×1024 resolution (8,000 training, 1,490 validation, 17,194 test). The test set comprises 16,413 anomalous and 781 normal images,

where pneumonia cases constitute the anomaly class.

2. Experiments on LiverCT

We performed a few pre-processing steps on the LiverCT [5, 13] benchmark. In this section, we introduce these techniques one by one and complete the Table 3

2.1. Data Preprocessing

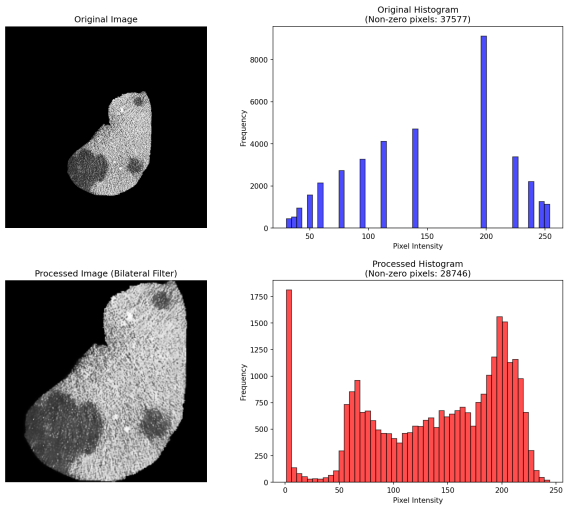

Figure 1. Original and processed images from LiverCT [5, 13] along with their pixel histograms.

First of all, the dimensions of images in the dataset are 512x512 pixels, and only small portion of the images have the Liver segments. Resizing these images to 224x224 pixels, which is the input size of our model, results in diminished liver sections. To resize without losing details of the region of interest (i.e. liver section) we used the following algorithm to get images resized to 224x224. Note that this process is fully automated and can be applied to any segmented liver images.

1. **ROI Identification:** For each 512×512 input image, we first identify the region containing the liver. This is achieved by computing a union bounding box that tightly encloses all non-zero pixels.
2. **ROI Cropping:** The image is cropped using the coordinates of the calculated bounding box, isolating the liver segment from the empty background.
3. **Canvas Preparation:** A new, black canvas of the target dimensions (224×224) is created to serve as the background for the final model input.
4. **Conditional Resizing and Placement:** The cropped liver ROI is placed onto the canvas using a size-dependent strategy:
   - **If the ROI is smaller than or equal to** 224×224**:** The cropped segment is pasted directly onto the center of the canvas without any resizing. This preserves the native resolution of the liver tissue.
   - **If the ROI is larger than** 224×224**:** The segment is resized to fit within the 224×224 frame while maintaining its original aspect ratio to prevent distortion. The resized ROI is then centered on the canvas.
5. **Final Input:** The resulting 224×224 image, with the liver segment prominently centered, is used as the input for the model.

Another issue with this dataset is that, due to constraints inherent to Computed Tomography imaging, it underwent several windowing and histogram equalization techniques [4, 6, 16]. As a result, these images can be out of distribution of standard datasets like ImageNet, on which our employed perceptual loss model is trained. To mitigate this, we apply a bilateral filter [26] to each processed 224×224 image prior to feeding it to the network.

The effect of the pre-processing is illustrated in Figure 1, where the ROI and anomalous regions are preserved, and the histogram of the image looks more natural.

Retraining and re-evaluating the model with this new preprocessing algorithm yieded the results in Row 2 of the Table 3. This result is mean and standard devation of evaluation of 5 different models trained with 5 different seed (42, 7, 13, 65, 91)

Table 3. Ablations on LiverCT Dataset.

	Version	AUROC
1	Main Config 1	54.1
2	+ Train & Eval with New Preprocessing	59.5 ± 1.27
3	+ Eval Perceptual Patch Sizes [16, 32, 56] $->$ [8, 16]	65.5 ± 1.96

2.2. New Evaluation Config

As shown in Figure 2, the new data preprocessing pipeline (column 2) has improved over the original config (column 1) regarding the quality of the anomaly map and made the anomalous region more visible. However, the predicted anomaly map still fails to capture the texture change in the anomalous region. Following the studies on visual perception [19, 20] that state smaller patch sizes are biased towards textures while larger patch sizes are biased towards shape, we changed the patch sizes used by perceptual model for anomaly score calculation from [16, 32, 56] to [8, 16]. As can be seen from the third column of Table 2, with this evaluation config anomaly maps capture texture changes on anomalous regions better. This reflects on the AUROC score of the 3rd row of the Table 3. The evaluation configuration that yields the best result on LiverCT is presented in Table 4, with the modified parts highlighted in bold. The training config is kept the same.

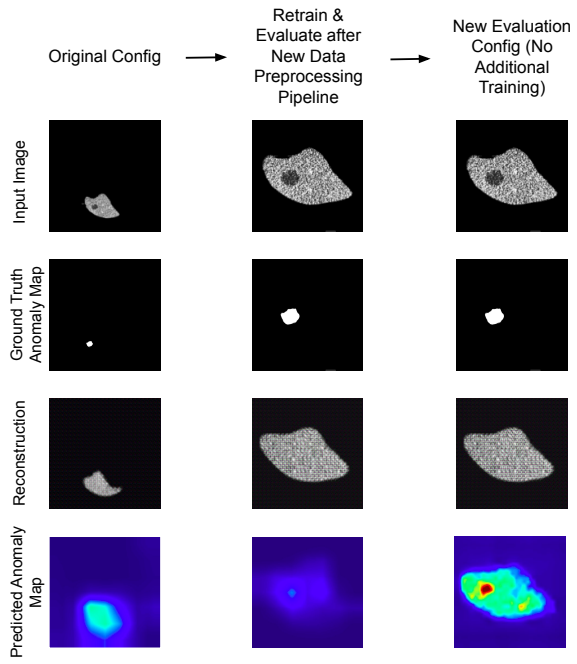

Figure 2. Effect of the modifications such as data preprocessing pipeline and evaluation config. We first avoid diminishing the anomalous region during resizing. Then configured perceptual loss to be more biased towards textual clues following insights from literature on visual perception.

Table 4. Best Evaluation Configuration on LiverCT.

Component	Parameter	Value
General	Batch Size	64
	Test Data Augmentation	None (only resize and normalize)
Perceptual Metric	Pre-trained Perceptual Model	MAE with ViT-Large Encoder
	Distance Metric	Cosine Distance
	Layers Used for Feature Extraction	From the 12th, 16th, and 20th transformer blocks
	Multi-Scale Input Patch Sizes	**8x8x, 16x16**
Image-Level Score Aggregation	Spatial Aggregation per Feature Map	Max
	Cross-Feature Map Aggregation	Mean
Pixel-Level Map Aggregation	Cross-Feature Map Aggregation	Mean

2.3. Results on LiverCT

We report the results obtained on the LiverCT dataset for our method, QFAE, and other state-of-the-art methods in Table 6. Our method, QFAE, achieves the second-highest performance with an AUROC score of 65.5%, being outperformed only by DRAEM [29] that scored 69.2%. The comparatively lower performance of all methods on this dataset can be attributed to its unique characteristics, specifically its sparse pixel distribution. As shown by the experiments above, the LiverCT dataset has a unique pixel distribution, being more sparse than the rest of the datasets. This pixel distribution differs from those of the natural images on which the encoders were trained.

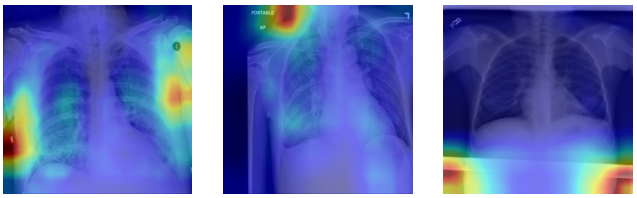

Figure 3. Optical characters and artifacts dominate the response from the anomalous region.

3. Different Aggregation for Chest RSNA

It is usual to see different optical characters and artifacts on Chest images. Their position varies. When these artifacts are present, they dominate the anomaly signals, and the anomaly score cannot be calculated properly with the aggregation method described in Main Eq. 4. Since their positions vary and are unpredictable, we were unable to devise a preprocessing algorithm.

To mitigate this problem, we decided to experiment with different aggregation methods on the validation split of the Chest RSNA dataset. As an alternative, we first tried taking the mean value in the anomaly map from each location, and then taking the maximum across different layers. We observed an increase in AUROC from 78.6% to 84.3% on the validation split. Therefore, we decided to keep this approach and reported an AUROC of 83.8% on test set as in main Table 5. The evaluation configuration that yields the best result on Chest RSNA is presented in 5, with the modified parts highlighted in bold. The training config is kept the same.

Table 5. Best Evaluation Configuration on Chest RSNA

Component	Parameter	Value
General	Batch Size	64
	Test Data Augmentation	None (only resize and normalize)
Perceptual Metric	Pre-trained Perceptual Model	MAE with ViT-Large Encoder
	Distance Metric	Cosine Distance
	Layers Used for Feature Extraction	From the 12th, 16th, and 20th transformer blocks
	Multi-Scale Input Patch Sizes	16x16, 32x32, 56x56
Image-Level Score Aggregation	Spatial Aggregation per Feature Map	**Mean**
	Cross-Feature Map Aggregation	**Max**
Pixel-Level Map Aggregation	Cross-Feature Map Aggregation	Mean

4. SOTA Results on Each Dataset

As shown in Table 7, we achieve the state-of-the-art performance in BraTS2021 [2, 3, 18], RESC [11] and RSNA [27] and second on LiverCT [5, 13].

References

[1] Samet Akcay, Amir Atapour-Abarghouei, and Toby P Breckon. Ganomaly: Semi-supervised anomaly detection via adversarial training. In *Asian conference on computer vision*, pages 622–637. Springer, 2018. 5

[2] Ujjwal Baid, Satyam Ghodasara, Suyash Mohan, Michel Bilello, Evan Calabrese, Errol Colak, Keyvan Farahani,

Table 6. Best Training Configuration for Brain MRI.

Component	Parameter	Value
General	Seed	42, 7, 13, 65, 91 (mean of 5 runs are reported)
	Image Resolution (Resize)	224x224
	Batch Size	64
	Epochs	300
	Device	CUDA
Encoder	Pre-trained Model	**ViT-L/14 + ViT-B/8**
	Pre-training Method	**DINOv2 + DINO**
	Frozen During Training	**True, True**
	Hidden States Used	Features from the 2nd and 4th to last blocks
	Final Projection In-Features	**1024, 768**
	Final Projection Out-Features	**768, 768**
Q-Former (Junction)	Number of Transformer Blocks	1
	Internal Dimension	768
	Output Dimension	768
	Number of Learnable Queries	784 (for 28x28 output patches)
	Attention Heads	8
	MLP Expansion Ratio	4.0
Decoder	Internal Dimension	768
	Depth (Number of Layers)	6
	Attention Heads	12
	Output Patch Size	8x8
	Number of Output Patches	28x28
	MLP Expansion Ratio	4.0
Optimization	Optimizer	Adam
	Learning Rate (Maximum)	8×10^{-5}
	Learning Rate Scheduler	OneCycleLR
Perceptual Loss	Pre-trained Perceptual Model	MAE with ViT-Large Encoder
	Distance Metric	Cosine Distance
	Layers Used for Feature Extraction	From the 16th and 20th transformer blocks
	Multi-Scale Input Patch Sizes	32x32, 56x56

Table 7. Anomaly detection performance (mean + std) on BraTS2021, Liver CT (BTCV + LiTs), RESC and RSNA. The results are reported for five repetitions of the experiment. *: denotes only three repetitions. The top results are reported in bold.

Methods	BraTS2021	Liver CT	RESC	RSNA
f-AnoGAN [25]	77.3 ± 0.18	58.4 ± 0.15	77.4 ± 0.85	55.6 ± 0.09
GANomaly [1]	74.8 ± 1.93	53.9 ± 2.36	52.6 ± 3.95	62.9 ± 0.65
DRAEM [29]	62.4 ± 9.03	$\mathbf{69.2 \pm 3.86}$	83.2 ± 8.21	67.7 ± 1.72
UTRAD [7]	82.9 ± 2.32	55.6 ± 5.96	89.4 ± 1.92	75.6 ± 1.24
DeepSVDD [23]	87.0 ± 0.66	53.3 ± 1.24	74.2 ± 1.29	64.5 ± 3.17
CutPaste [15]	78.8 ± 0.67	58.6 ± 4.2	90.2 ± 0.61	82.6 ± 1.22
SimpleNet [17]	82.5 ± 3.34	N/A	76.2 ± 7.46	69.1 ± 1.27
MKD [24]	81.5 ± 0.36	60.4 ± 1.61	89.0 ± 0.25	82.0 ± 0.12
RD4AD [9]	89.5 ± 0.91	60.0 ± 1.4	87.8 ± 0.87	67.6 ± 1.11
STFPM [28]	83.0 ± 0.67	61.6 ± 1.7	84.8 ± 0.50	72.9 ± 1.96
PaDiM [8]	79.0 ± 0.38	50.7 ± 0.5	75.9 ± 0.54	77.5 ± 1.87
PatchCore [21]	91.7 ± 0.36	60.4 ± 0.82	91.6 ± 0.10	76.1 ± 0.67
CFA [14]	84.4 ± 0.87	61.9 ± 1.16	69.9 ± 0.26	66.8 ± 0.23
CFLOW [10]	74.8 ± 5.32	49.9 ± 4.67	75.0 ± 5.81	71.5 ± 1.49
CS-Flow [22]	90.9 ± 0.83	59.4 ± 0.52	87.3 ± 0.58	83.2 ± 0.46
P-VQ* [12]	94.3 ± 0.23	60.6 ± 0.62	89.0 ± 0.48	79.2 ± 0.04
QFAE (ours)	$\mathbf{94.3 \pm 0.18}$	65.5 ± 1.96	$\mathbf{91.8 \pm 0.55}$	$\mathbf{83.8 \pm 0.46}$

Jayashree Kalpathy-Cramer, Felipe C Kitamura, Sarthak Pati, et al. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv preprint arXiv:2107.02314*, 2021. 2, 4

[3] 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. 2, 4

[4] Jinan Bao, Hanshi Sun, Hanqiu Deng, Yinsheng He, Zhaoxiang Zhang, and Xingyu Li. Bmad: Benchmarks for medical anomaly detection. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4042–4053, 2024. 2, 3

[5] Patrick Bilic, Patrick Ferdinand Christ, Eugene Vorontsov, Grzegorz Chlebus, Hao Chen, Qi Dou, Chi-Wing Fu, Xiao Han, Pheng-Ann Heng, Jürgen Hesser, et al. The liver tumor segmentation benchmark (lits). *arXiv preprint arXiv:1901.04056*, 2019. 2, 3, 4

[6] Patrick Bilic, Patrick Christ, Hongwei Bran Li, Eugene Vorontsov, Avi Ben-Cohen, Georgios Kaissis, Adi Szeskin, Colin Jacobs, Gabriel Efrain Humpire Mamani, Gabriel Chartrand, et al. The liver tumor segmentation benchmark (lits). *Medical Image Analysis*, 84:102680, 2023. 3

[7] Liyang Chen, Zhiyuan You, Nian Zhang, Juntong Xi, and Xinyi Le. Utrad: Anomaly detection and localization with u-transformer. *Neural Networks*, 147:53–62, 2022. 5

[8] Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. Padim: a patch distribution modeling framework for anomaly detection and localization. In *International Conference on Pattern Recognition*, pages 475–489. Springer, 2021. 5

[9] Hanqiu Deng and Xingyu Li. Anomaly detection via reverse distillation from one-class embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9737–9746, 2022. 5

[10] Denis Gudovskiy, Shun Ishizaka, and Kazuki Kozuka. Cflow-ad: Real-time unsupervised anomaly detection with localization via conditional normalizing flows. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 98–107, 2022. 5

[11] Junjie Hu, Yuanyuan Chen, and Zhang Yi. Automated segmentation of macular edema in oct using deep neural networks. *Medical image analysis*, 55:216–227, 2019. 2, 4

[12] Taejune Kim, Yun-Gyoo Lee, Inho Jeong, Soo-Youn Ham, and Simon S. Woo. Patch-wise vector quantization for unsupervised medical anomaly detection. *Pattern Recognition Letters*, 184:205–211, 2024. 5

[13] Bennett Landman, Zhoubing Xu, J Igelsias, Martin Styner, T Langerak, and Arno Klein. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In *Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge*, page 12, 2015. 2, 3, 4

[14] Sungwook Lee, Seunghyun Lee, and Byung Cheol Song. Cfa: Coupled-hypersphere-based feature adaptation for target-oriented anomaly localization. *IEEE Access*, 10: 78446–78454, 2022. 5

[15] Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. Cutpaste: Self-supervised learning for anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9664–9674, 2021. 5

[16] He Li, Yutaro Iwamoto, Xianhua Han, Lanfen Lin, Hongjie Hu, and Yen-Wei Chen. An accurate unsupervised liver lesion detection method using pseudo-lesions. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 214–223. Springer, 2022. 3

[17] Zhikang Liu, Yiming Zhou, Yuansheng Xu, and Zilei Wang. Simplenet: A simple network for image anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20402–20411, 2023. 5

[18] 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. 2, 4

[19] Muzammal Naseer, Kanchana Ranasinghe, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers, 2021. 3

[20] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks?, 2022. 3

[21] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14318–14328, 2022. 5

[22] Marco Rudolph, Tom Wehrbein, Bodo Rosenhahn, and Bastian Wandt. Fully convolutional cross-scale-flows for image-based defect detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1088–1097, 2022. 5

[23] Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *International conference on machine learning*, pages 4393–4402. PMLR, 2018. 5

[24] Mohammadreza Salehi, Niousha Sadjadi, Soroosh Baselizadeh, Mohammad H Rohban, and Hamid R Rabiee. Multiresolution knowledge distillation for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14902–14912, 2021. 5

[25] Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Georg Langs, and Ursula Schmidt-Erfurth. f-anogan: Fast unsupervised anomaly detection with generative adversarial networks. *Medical image analysis*, 54:30–44, 2019. 5

[26] C. Tomasi and R. Manduchi. Bilateral filtering for gray and color images. In *Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271)*, pages 839–846, 1998. 3

[27] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017. 2, 4

[28] Shinji Yamada and Kazuhiro Hotta. Reconstruction student with attention for student-teacher pyramid matching. *arXiv preprint arXiv:2111.15376*, 2021. 5

[29] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Draem-a discriminatively trained reconstruction embedding for surface anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8330–8339, 2021. 4, 5