

This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Supplementary Material

1 Additional Experimental Results

1.1 Ablation Studies

All ablation studies are conducted on five cases from the FLARE22 dataset. The selected hyperparameters are adopted in all subsequent experiments.

Diffusion noise level. We evaluate the effect of the diffusion noise level t (Fig. 1(a)). The best performance is obtained at $t = 0.2$, corresponding to the highest Dice score. Performance remains relatively stable across the tested range, indicating low sensitivity to this parameter.

Feature layer selection. Figs. 1(b) and 1(c) show the results for diffusion and self-attention feature layer selection. For diffusion features, the Up4 decoder block achieves the best performance. For self-attention features, the Down3 block provides the most effective representation.

Recursive depth. Fig. 1(d) illustrates the effect of recursive N-Cuts depth. Increasing the recursion depth generally improves segmentation performance up to a certain point, with depth 5 yielding the best result.

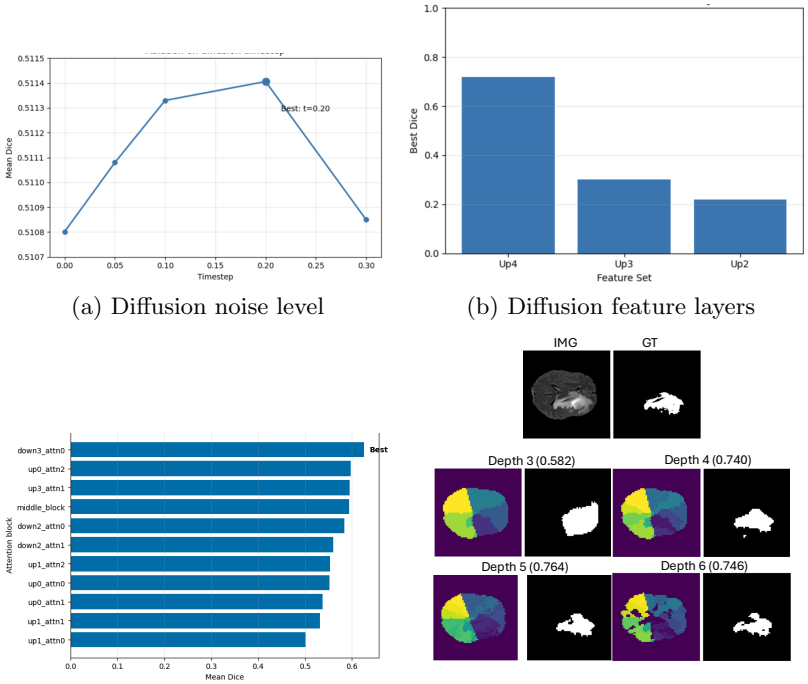
1.2 FLARE22 Results under 2D Evaluation

To complement the 3D evaluation reported in the main paper, Table 1 presents per-organ segmentation results on the FLARE22 dataset under the 2D evaluation protocol. Consistent with the observations from the 3D evaluation, the proposed framework achieves the highest average Dice and IoU scores across organs. These results further support the benefit of integrating complementary feature representations for unsupervised abdominal CT segmentation.

2 Detailed Hyperparameter and Preprocessing Settings

2.1 Preprocessing

All images are preprocessed following the MAISI protocol [2]. Volumes are first reoriented to the RAS coordinate system. For BraTS2020, MRI intensities are normalised using non-zero voxels with percentile scaling (0–99.5%) and clipped to the range $[0, 1]$. For FLARE22, CT intensities are clipped to the range $[-160, 300]$ HU and linearly scaled to $[0, 1]$. All images are resized to fixed spatial dimensions of $256 \times 256 \times 256$ for BraTS2020 and $256 \times 256 \times 128$ for FLARE22. Trilinear interpolation is used for image volumes, while nearest-neighbour interpolation is used for label maps during evaluation.



(a) Diffusion noise level (b) Diffusion feature layers

(c) Self-attention feature layers (d) Recursive N-Cuts depth

Fig. 1. Ablation studies on FLARE22. (a) Effect of diffusion noise level t. (b) Diffusion U-Net feature layer selection. (c) Self-attention layer selection. (d) Effect of recursive N-Cuts depth. The selected configuration ($t = 0.2$, Up4, Down3, and depth 5) is used in all subsequent experiments.

2.2 Feature Extraction and Graph Construction

The pretrained MAISI VAE encoder is used to generate latent representations with four channels. Diffusion features are extracted from the fourth decoder upsampling block (Up4) of the pretrained diffusion U-Net using a single forward pass at a fixed diffusion noise level of $t = 0.2$. Self-attention features are extracted from the third downsampling block (Down3). The attention maps are averaged across attention heads, symmetrised, transformed into spatial feature maps, and reduced using PCA. The PCA dimension is set to 512 for BraTS2020 and 64 for FLARE22.

For graph construction, all feature representations are resampled to a common spatial resolution of $64 \times 64 \times 64$. Spatial coordinates are incorporated into all feature representations with a weighting factor of 0.01. Affinity matrices are computed using cosine similarity and fused using weights $(w_I, w_A, w_B, w_C) = (1.0, 0.5, 1.0, 0.2)$, corresponding to image intensity, VAE latent representation, diffusion features, and self-attention features, respectively.

Table 1. Per-organ segmentation performance on the FLARE22 dataset under 2D evaluation. Bold values indicate the best results for each metric.

	Ours			Intensity			DiffCut			CUTS		
Organ	Dice ↑	IoU ↑	HD ↓	Dice ↑	IoU ↑	HD ↓	Dice ↑	IoU ↑	HD ↓	Dice ↑	IoU ↑	HD ↓
Liver	0.516	0.399	86.389	0.531	0.410	**73.610**	**0.602**	**0.497**	84.784	0.396	0.295	143.207
Right Kidney	0.285	0.192	44.139	**0.304**	**0.210**	**41.733**	0.139	0.090	97.472	0.050	0.028	146.993
Spleen	**0.443**	**0.328**	**39.004**	0.394	0.283	45.802	0.069	0.047	102.171	0.040	0.022	155.933
Pancreas	**0.114**	**0.070**	31.580	0.101	0.064	**31.142**	0.053	0.031	68.260	0.022	0.012	100.607
Aorta	**0.091**	**0.056**	**28.312**	0.046	0.028	28.363	0.063	0.042	51.269	0.002	0.001	172.139
IVC	**0.040**	**0.025**	**18.761**	0.021	0.013	22.847	0.010	0.007	58.538	0.002	0.001	149.547
Right Adrenal	**0.054**	**0.031**	24.620	0.042	0.025	**20.549**	0.001	0.001	87.738	0.001	0.001	121.583
Left Adrenal	0.028	0.017	**16.220**	**0.041**	**0.025**	24.658	0.000	0.000	62.524	0.002	0.001	128.275
Gallbladder	**0.161**	**0.104**	**29.486**	0.107	0.068	31.871	0.038	0.027	93.429	0.026	0.015	140.974
Esophagus	**0.034**	**0.020**	40.062	0.014	0.009	**27.421**	0.022	0.013	60.214	0.001	0.000	175.955
Stomach	0.292	0.187	**46.935**	0.284	0.181	46.494	**0.311**	**0.212**	72.136	0.059	0.032	182.643
Duodenum	**0.118**	**0.070**	38.702	0.101	0.061	**37.647**	0.021	0.012	64.081	0.010	0.005	127.463
Left Kidney	0.380	0.267	**36.499**	**0.383**	**0.276**	42.129	0.143	0.094	89.943	0.046	0.025	127.577
Average	**0.197**	**0.136**	36.978	0.182	0.127	**36.482**	0.113	0.083	76.351	0.051	0.034	144.069

2.3 Runtime

The average processing time, including both feature extraction and recursive normalized cuts, is approximately 10 minutes per volume for BraTS2020 and 20 minutes per volume for FLARE22. Feature extraction is performed on an NVIDIA RTX A6000 GPU, while recursive normalized cuts are executed on the CPU. The longer runtime on FLARE22 is primarily due to the increased complexity of multi-organ segmentation and the larger foreground region considered during graph construction.

2.4 Baseline Methods

All baseline methods follow the same preprocessing pipeline and evaluation protocol as our approach.

Intensity-only baseline. As a simple baseline, we apply 3D normalized cuts directly on the input intensity augmented with spatial coordinates (x, y, z). This baseline does not rely on learned features and serves to evaluate the effectiveness of classical graph-based segmentation.

DiffCut [1]. A diffusion based unsupervised segmentation method built on SSD-1B, a distilled version of Stable Diffusion XL. It performs segmentation in 2D using a graph-based formulation, in contrast to our 3D approach. Since it operates on 2D images, it is applied slice-wise to each volume, and the resulting slice-wise predictions are stacked to form a 3D segmentation. We perform an ablation study to determine suitable hyperparameters and adopt $\tau = 0.80$, $\alpha = 10$, and diffusion timestep $t = 150$ in our experiments. To ensure consistent labeling across slices, the stacked predictions are relabeled using Hungarian matching [3] based on Dice overlap between segments in adjacent slices.

CUTS [4]. A 2D unsupervised segmentation framework based on contrastive representation learning and diffusion condensation. We train the model on their

brain tumour dataset, following the original protocol consisting of 204 cases, to obtain a pretrained encoder. During inference, the pretrained encoder is applied in a zero-shot manner to both datasets without any fine-tuning. Feature embeddings are extracted per slice and grouped using diffusion condensation, producing a hierarchy of segmentation maps at different levels of granularity. For BraTS2020, which is evaluated as a binary segmentation task, the final clustering level is selected. For FLARE22, which involves multi-organ segmentation, an intermediate clustering level is used to preserve finer anatomical structures. The resulting slice-wise predictions are stacked into 3D volumes and processed using the same relabeling pipeline as DiffCut.

2.5 Cluster-to-Ground-Truth Matching

Different matching strategies are adopted for the two datasets due to their distinct segmentation tasks. For BraTS2020, all tumour subregions are merged into a single foreground class. A predicted cluster is retained if more than 20% of its voxels overlap with the lesion region, i.e., the intersection-to-cluster ratio exceeds 0.20. All selected clusters are subsequently merged to form the final binary segmentation. This procedure is applied in both 2D and 3D evaluation settings.

For FLARE22, which requires multi-organ segmentation, background clusters are first removed using a purity threshold of 85%. The remaining clusters are separated using connected component analysis with 6-connectivity, and components smaller than 200 voxels are discarded. Each remaining component is then assigned to the organ with the highest Dice overlap using a greedy matching strategy. Multiple components may be assigned to the same organ, allowing fragmented predictions to contribute to the final segmentation. The matched components are merged to obtain the final organ-wise predictions for evaluation.

References

1. Couairon, P., Shukor, M., Haugeard, J.E., Cord, M., Thome, N.: Diffcut: Catalyzing zero-shot semantic segmentation with diffusion features and recursive normalized cut. Advances in Neural Information Processing Systems **37**, 13548–13578 (2024)
2. Guo, P., Zhao, C., Yang, D., Xu, Z., Nath, V., Tang, Y., Simon, B., Belue, M., Harmon, S., Turkbey, B., et al.: Maisi: Medical ai for synthetic imaging. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 4430–4441. IEEE (2025)
3. Kuhn, H.W.: The hungarian method for the assignment problem. Naval research logistics quarterly **2**(1-2), 83–97 (1955)
4. Liu, C., Amodio, M., Shen, L.L., Gao, F., Avesta, A., Aneja, S., Wang, J.C., Del Priore, L.V., Krishnaswamy, S.: Cuts: A deep learning and topological framework for multigranular unsupervised medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 155–165. Springer (2024)