Proceedings of the 4th International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/2026.37692

Functional Roles of Diffusion Models in Medical Image Segmentation: A Focused Narrative Review

Chenyang Li

Computer Software Development, University of Electronic Science and Technology of China, Chengdu, China
18955301787@163.com

Abstract. Diffusion-based medical image segmentation has expanded quickly, but the term now covers methods that solve different problems. Existing reviews often organize the field by architecture or application, which can blur the distinction between a model that generates a mask, one that supplies a prior, one that repairs an existing prediction, and one that models disagreement. This focused review uses those four operational roles as its organizing principle. Direct generators can recover weak boundaries and produce several plausible contours, although repeated sampling remains costly. Recent latent and categorical designs reduce that burden, while refinement methods avoid reconstructing the whole mask from noise. Representation-oriented approaches are promising but still supported by less direct medical-segmentation evidence. Uncertainty-oriented methods are most informative when sample variation is connected to calibration, error, or expert disagreement rather than treated as confidence by default. Across the literature, fair comparison is hindered by different datasets, preprocessing pipelines, sampling schedules, ensemble rules, and hardware. The evidence suggests that diffusion should be judged according to the role it performs. Progress will depend on efficient 3D inference, external validation, anatomically credible outputs, reproducible reporting, and prospective evidence of clinical benefit.

Keywords: diffusion models, medical image segmentation, latent diffusion, mask refinement, uncertainty quantification

1. Introduction

Medical image segmentation turns an image or volume into a spatial description of anatomy or disease. Those contours are used to measure lesions, plan treatment, guide procedures, and assess change over time. U-Net established the encoder-decoder pattern that still underlies much of biomedical segmentation and showed that useful localization could be learned from comparatively small annotated datasets [1]. Strong discriminative models have since improved context modeling and multiscale feature extraction, but they do not remove several difficulties that are intrinsic to the task. Small targets can occupy only a few pixels, boundaries may fade into surrounding tissue, artifacts can obscure local evidence, and more than one contour may be clinically defensible.

These difficulties explain why diffusion became attractive for segmentation rather than being adopted simply because it was successful in image generation. A denoising diffusion probabilistic

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.37692

model (DDPM) learns to reverse a gradual corruption process [2]. When the corrupted variable is a label map and the image provides the condition, the reverse trajectory offers a way to build or revise a structured mask step by step. Repeated sampling can also expose alternative solutions that a single deterministic output cannot show. The same mechanism, however, creates a practical tension: more reverse steps and more samples may enrich the output distribution, but they also increase inference cost. Whether diffusion is useful therefore depends on what the diffusion component is asked to do, not merely on whether it is present.

Previous reviews provide important background but do not answer that question in the same way. Yang et al. surveyed diffusion methods across computer vision, so medical segmentation appears as only one application within a much wider methodological landscape [3]. Kazerouni et al. reviewed diffusion throughout medical imaging [4], while Shi et al. organized medical image computing into several task groups [5]. Their breadth is valuable for orientation, but it also means that segmentation-specific differences in the function of diffusion receive limited emphasis. The 2026 literature narrows the scope from different directions. Yaseen et al. provide a dedicated account of diffusion-based medical segmentation [6], yet the operational role of the diffusion component is not the main organizing axis. Azad et al. focus on diagnostic imaging, clinical integration, and explainability across multiple tasks [7], whereas Sabanci et al. compare CNNs, transformers, foundation models, and diffusion systems for segmentation-oriented model selection [8]. These studies are useful, but they leave an interpretive gap: a full mask generator, a prior-guided refiner, and a model of annotator disagreement may all be grouped under 'diffusion segmentation' even though the appropriate baseline and evaluation target are different.

The motivation for this review is to make those differences explicit. If methods with different jobs are compared by one headline Dice score, the comparison can be misleading: a refiner partly inherits the strength of its first-stage model, an uncertainty model should be judged on calibration or agreement as well as overlap, and a direct generator must justify the cost of repeated denoising. The four roles used here are intentionally overlapping, because one method can serve more than one function. Organizing the literature by function provides a clearer link between mechanism, evidence, computational cost, and the question of whether diffusion adds value beyond a simpler discriminative alternative.

2. Technical foundations

A DDPM consists of a forward Markov chain that corrupts a clean sample and a learned reverse chain that reconstructs it [2]. During training, a U-Net-like denoiser typically receives a noisy sample together with a time step and estimates the added noise. Improved DDPM showed that reverse variance and sampling schedules can be redesigned to reduce the number of network evaluations without sacrificing sample quality [9]. Latent diffusion changes where denoising takes place: the reverse process operates on a compressed representation rather than the original image or label grid, reducing memory use and making conditioning easier to manage [10].

For segmentation, the diffusion state has taken several forms, including continuous masks, one-hot channels, learned shape latents, binary variables, and categorical distributions. Conditioning is equally varied: the medical image may be concatenated with the noisy mask, processed by a separate encoder, injected at several scales, or connected through attention. Gaussian corruption is convenient but does not naturally match discrete labels. HiDiff instead uses Bernoulli transitions for binary masks [11], while CRAC-DM introduces class-aware categorical transitions for multi-class surgical scenes [12]. Both choices keep the intermediate diffusion state closer to the semantics of a segmentation label.

DOI: 10.54254/2755-2721/2026.37692

Diffusion can also be useful before a final mask is produced. Perception-prioritized training showed that intermediate noise levels encourage recovery of semantic structure rather than local texture alone [13]. This observation helps explain why later work reused attention maps, hidden features, synthetic data, and fixed generative priors. The transfer problem should not be overlooked, however. A representation learned from natural images may encode object-level semantics while missing subtle intensity changes, scanner artifacts, or thin structures that matter in MRI, CT, ultrasound, pathology, and fundus imaging.

Four design decisions are especially important when a diffusion segmentation result is interpreted: the space in which diffusion operates, the route by which the image conditions the model, the state from which reverse sampling starts, and the number of steps or samples used. A model that starts from random noise is solving a different problem from one that starts near a preliminary mask. Likewise, an ensemble of sampled masks should not be compared casually with one deterministic prediction. Accuracy is meaningful only when it is read alongside boundary error, latency, memory, step count, sample count, and the rule used to combine outputs.

3. Functional roles of diffusion in segmentation

The role assignments used in this review are summarized in Table 1. Each method is given a primary role and, where it changes interpretation, a secondary role. The categories overlap deliberately. More important than the label itself is the comparison that follows from it. A refiner should be tested against the same first-stage segmentor before correction; an acceleration strategy against the same model under another sampling schedule; and an uncertainty method against calibration or annotator-distribution measures. Selected quantitative results from the primary studies are collected in Table 2. They are included to illustrate role-specific evidence rather than to create a cross-paper ranking.

This distinction also changes what counts as failure. A direct generator may lose a small structure during denoising or depend too heavily on weak conditioning. A useful representation can still transfer poorly to a particular modality. A refiner may appear strong largely because its base segmentor is already strong, while a stochastic model can produce varied but badly calibrated masks. Reading Tables 1 and 2 together is therefore more informative than treating every reported Dice score as evidence for the same claim.

Table 1. Functional roles of representative methods. Primary and secondary roles are not mutually exclusive

Method	Primary role	Secondary role	Evidence domain
SegDiff [14]	Mask generator	Implicit ensemble	Natural images; MoNuSeg
MedSegDiff [15]	Mask generator	Implicit ensemble	Fundus, MRI, ultrasound
MedSegDiff-V2 [16]	Mask generator	Anchor prior; diversity	Five medical modalities
Diff-UNet [17]	Embedded 3D generator	Uncertainty-weighted fusion	MRI and CT volumes
FCFDiff-Net [18]	Embedded 3D generator	Conditional feature fusion	Brain MRI
LDSeg [19]	Latent mask generator	Acceleration; robustness	Echo, pathology, knee MRI
CRAC-DM [12]	Categorical mask generator	Acceleration; local uncertainty	Surgical video
HiDiff [11]	Mask refiner	Discrete generative modeling	CT, MRI, endoscopy, fundus
PD-DDPM [20]	Mask refiner	Acceleration; uncertainty	Brain MRI

DOI: 10.54254/2755-2721/2026.37692

Table 1. (continued)

DiffRect [21]	Label rectifier	Semi-supervised learning	Cardiac and prostate MRI
CIMD [22]	Expert-disagreement model	Mask generation	CT, ultrasound, MRI
DiffSeg [23]	Representation/prior	Training-free grouping	Natural images
Graikos et al. [24]	Task prior	Constraint-guided transfer	Remote sensing

Note: The table describes the function of diffusion rather than methodological quality or a universal performance ranking.

Table 2. Selected results reproduced from the primary studies

Method	Dataset / setting	Reported evidence
MedSegDiff-L [15]	BraTS-2021	Dice 89.9%; HD95 8.72 mm
MedSegDiff-V2 [16]	AMOS / BTCV	Average Dice 90.1% / 89.5%
Diff-UNet [17]	BraTS-2020 / BTCV	Dice 85.35% / 83.75%; BTCV HD95 8.115 mm
FCFDiff-Net [18]	BraTS-2021	Dice 89.9%; HD95 1.858 mm
LDSeg [19]	Echo, 2 DDIM steps	Combined Dice 0.92; 0.34 s per image
CRAC-DM [12]	CholecSeg8k, 10 steps	mIoU 84.79%; mDice 91.37%; 30 FPS
HiDiff [11]	Synapse	Dice 83.75% to 84.94%; HD95 16.45 to 13.42 mm
PD-DDPM [20]	WMH, start step 300	Dice 0.812; HD95 3.494 mm
DiffRect [21]	ACDC, 1% labels	Dice 82.40%; +4.60 percentage points over cited prior best
CIMD [22]	LIDC-IDRI	GED 0.321; CI 0.759; maximum Dice 0.915

Note: Results are not directly comparable across papers because datasets, targets, preprocessing, splits, sampling schedules, ensemble sizes, and hardware differ.

3.1. Direct and embedded mask generators

For direct generation, the diffusion process is responsible for the final segmentation. Recovery may begin from random noise, a strongly corrupted label, or a low-dimensional latent, with the image guiding the reverse trajectory. The appeal is straightforward: global position and shape can emerge before local boundaries are restored. The drawback is equally clear. The denoiser is evaluated repeatedly, and the cost increases again when several masks are sampled and combined.

SegDiff provided an early demonstration of this idea [14]. Separate encoders process the image and the current mask estimate, their features are combined within a U-Net, and several sampled masks are merged. On MoNuSeg, the reported Dice and mIoU were 81.59% and 69.00%. The experiments established feasibility, although they covered natural, aerial, and microscopy images rather than a broad set of clinical modalities.

MedSegDiff moved conditional diffusion toward optic-cup, brain-tumor, and thyroid-nodule segmentation [15]. Dynamic Conditional Encoding combines image and noisy-mask features inside the network, while a Fourier-domain Feature Frequency Parser suppresses components associated with noise. On BraTS-2021, MedSegDiff-L reported 89.9% Dice and 8.72 mm HD95; the EnsemDiff baseline reproduced in the same table reported 88.7% Dice. The tabulated values therefore support a 1.2-percentage-point difference, despite a larger improvement stated elsewhere in the source. The appendix also reported 19.76 s per image for the large model, compared with 0.03 s for nnU-Net, which illustrates the cost of classical iterative inference.

DOI: 10.54254/2755-2721/2026.37692

MedSegDiff-V2 strengthens the exchange between the image and the evolving mask through an Anchor Condition, Uncertain Spatial Attention, and a Spectrum-Space Transformer [16]. Average Dice was reported as 90.1% on AMOS and 89.5% on BTCV; the reported BraTS comparison gave 90.8% Dice and 7.53 mm HD95. The anchor supplies a semantic reference, but diffusion still produces the final mask. It is therefore more accurately treated as a direct generator with a secondary prior role than as an independent post-processing stage.

Volumetric diffusion is represented by Diff-UNet and FCFDiff-Net. Diff-UNet combines one-hot label diffusion, a separate feature encoder, and Step-Uncertainty Fusion across DDIM steps [17]. It reported 85.35% average Dice on BraTS-2020 and 83.75% Dice with 8.115 mm HD95 on BTCV; uncertainty-weighted fusion increased the BraTS average from 84.76% to 85.35%. FCFDiff-Net uses full-conditional feature embedding and multi-head attention residual fusion [18]. Its reported Dice/HD95 results were 85.4%/2.356 mm on BraTS-2020 and 89.9%/1.858 mm on BraTS-2021.

LDSeg, formally published in 2026, learns low-dimensional spaces for both images and labels [19]. The label encoder maps discrete masks onto a standardized latent shape manifold, and the encoder, denoiser, and decoder are trained jointly. Reported combined Dice scores were 0.92 for echocardiography, 0.91 for GlaS histopathology, and 0.96 for knee MRI. With two DDIM steps, the corresponding per-image times were 0.34 s, 0.30 s, and 0.49 s, compared with 91.23 s, 80.37 s, and 132.36 s in the authors' 1,000-step setting. These are within-method measurements on an NVIDIA A100, not a universal speed ranking.

CRAC-DM addresses multi-class labels through class-aware categorical diffusion [12]. A transition matrix derived from pretrained visual features biases corruption toward semantically related classes. Step-skipping shortens the reverse chain, and entropy-triggered test-time augmentation revisits uncertain regions. On CholecSeg8k, the 10-step model reported 84.79% mIoU and 91.37% mDice, whereas CCDM achieved 81.81% mIoU. In the authors' implementation, reducing the chain from 250 to 10 steps raised throughput from 2 to 30 FPS and reduced the estimated diffusion cost from 8.41 to 0.34 TFLOPs. On EndoVis18, mIoU was 48.74%, compared with 47.28% for CCDM.

Taken together, these methods show that direct generation is no longer synonymous with a long pixel-space Gaussian chain. Latent and categorical designs can make the reverse process much shorter. That does not make the cost irrelevant. Direct generation is easiest to justify when the task genuinely benefits from iterative reconstruction, class-aware transitions, robustness to difficult boundaries, or multiple plausible outputs. On routine structures with little ambiguity, a strong feed-forward segmentor may still be the more sensible choice.

The evidence also resists simple ranking. Studies use different public datasets, preprocessing pipelines, target definitions, sampling schedules, and ensemble rules. Some headline results come from one sample, others from several fused predictions. For that reason, single-sample and ensemble performance should be separated, and the same backbone should be tested with and without the diffusion objective. Authors should also report where additional steps or samples stop producing a meaningful gain. Without those controls, extra accuracy cannot be cleanly attributed to diffusion rather than to added capacity or computation.

3.2. Representation and task-prior provision

A second use of diffusion is more indirect. Instead of producing the final mask, the model contributes an attention pattern, an internal representation, synthetic training data, or a fixed prior that guides another predictor. This can avoid a complete reverse chain at inference, but the supporting evidence is less direct in medical segmentation than it is for generation and refinement.

DOI: 10.54254/2755-2721/2026.37692

DiffSeg reuses self-attention maps from Stable Diffusion and merges them according to spatial similarity [23]. The method is prompt-free and requires no additional training; it reported 72.5% pixel accuracy and 43.6% mIoU on COCO-Stuff-27. Graikos et al. instead treated an independently trained DDPM as a plug-and-play prior under differentiable constraints [24]. In an EnviroAtlas transfer experiment, a model trained on inferred labels reached 79.8% accuracy and 46.4% mIoU in Durham. Both studies demonstrate mechanisms outside medical segmentation rather than direct superiority on clinical tasks.

The medical examples in this category mainly concern synthetic pretraining. Khader et al. generated 3D CT and MRI data and reported an increase in breast-segmentation Dice from 0.91 to 0.95 in a limited-data experiment [25]. Med-DDPM used semantic masks to control 3D brain MRI synthesis [26]. A tumor segmentor trained on generated images reached 0.6207 Dice, compared with 0.6531 on real images; combining real and synthetic data increased the value to 0.6675. These findings support diffusion as a data prior, but they do not show that a frozen natural-image diffusion backbone is the best representation for medical segmentation.

The limitation is transfer. Natural-image attention may organize large objects well while overlooking weak tissue contrast, scanner-specific artifacts, or thin anatomy. Synthetic pretraining can improve data efficiency but can also reproduce the generator's anatomical mistakes or dataset bias. Stronger evidence would require matched comparisons with domain-specific self-supervision and medical foundation models under the same annotation budget, backbone size, and external test set. Until such comparisons are routine, representation and prior reuse should be treated as a promising supporting role rather than a well-established replacement for medical segmentation features.

3.3. Prediction refinement and label rectification

Refinement methods start much closer to the desired answer. Their input is a prediction, pseudo-label, or label feature that already resembles the target, and diffusion is used to repair, regularize, or redistribute that estimate. This changes the evaluation logic: the relevant baseline is the same model before refinement, not an unrelated segmentation architecture.

PD-DDPM starts from a mask produced by a separately trained pre-segmentation network [20]. That mask is corrupted only to an intermediate point and is then recovered along a shorter reverse trajectory. On WMH, a start step of 300 produced Dice 0.812 and HD95 3.494 mm, compared with 0.796 and 4.179 mm for vanilla DDPM. Because the quality of the first-stage prediction matters, the method is best understood as both acceleration and refinement.

HiDiff makes the hybrid structure explicit [11]. A discriminative segmentor supplies a prior mask, and Bernoulli diffusion refines it. Time-dependent binarization reduces computation, while a binary cross-transformer exchanges discriminative and generative features. On Synapse, refinement increased average Dice from 83.75% to 84.94% and reduced HD95 from 16.45 to 13.42 mm. Dice rose from 95.32% to 96.29% on Kvasir-SEG and from 80.24% to 81.72% on DRIVE, where HD95 fell from 8.88 to 4.33. The study used ten sampling steps and reported a 22-fold speed-up for the binarized refiner relative to its real-valued counterpart.

DiffRect moves correction into semi-supervised training [21]. Label Context Calibration first revises category relations in pseudo-labels, after which Latent Feature Rectification performs coarse-to-fine and fine-to-precise transport. With 1% labeled ACDC scans, the method reported 82.40% Dice, 4.60 percentage points above the previous best result identified by the authors. Its main contribution is improved supervision during training rather than a separate post-processing stage used at the clinic.

DOI: 10.54254/2755-2721/2026.37692

This hybrid arrangement is attractive because a fast segmentor can handle routine cases and diffusion can be reserved for correction. The additional module still has to earn its complexity. A fair study should compare it with cheaper alternatives such as a second decoder, boundary-aware loss, test-time augmentation, or conventional post-processing. The issue is especially important in semi-supervised learning, where better pseudo-labels can also come from stronger augmentation, thresholding, or consistency training. Ablations need to isolate the correction mechanism rather than compare two pipelines that differ in several places at once.

3.4. Uncertainty and expert-disagreement modeling

Stochastic sampling makes uncertainty analysis possible, but it also creates an easy source of over-interpretation. Several different masks may indicate genuine ambiguity, model instability, or simply a poorly chosen noise schedule. Variation becomes useful only when it is linked to error, independent raters, calibration, or a decision in which review or abstention matters.

CIMD is designed around observed disagreement between experts [22]. Its ambiguity-modeling and ambiguity-controlling networks are trained so that generated masks resemble the annotation distribution. On LIDC-IDRI, it reported generalized energy distance 0.321, Collective Insight 0.759, and maximum Dice 0.915; Probabilistic U-Net obtained 0.353, 0.731, and 0.892. Experiments on bone ultrasound and multiple-sclerosis MRI support the interpretation that CIMD models structured inter-rater variation rather than functioning as a simple ensemble.

Uncertainty is a secondary function in several other systems. PD-DDPM found that predictive variance changed with the selected starting step [20]. Diff-UNet used uncertainty from individual denoising steps to weight fusion and gained 0.59 Dice points over simple fusion [17]. CRAC-DM used entropy to decide where local test-time augmentation was worth the added cost [12]. These mechanisms identify unstable regions, but they do not by themselves establish calibration or cleanly separate aleatoric from epistemic uncertainty.

A useful counterpoint comes from deep evidential fusion [27]. MMEF-nnFormer models modality reliability through Dempster-Shafer evidence and reported mean Dice 0.893 with expected calibration error 0.052 on BraTS-2021. Diffusion studies that make clinical confidence claims need similarly explicit tests, including calibration error, risk-coverage or error-retention curves, sensitivity to sample count, and agreement with held-out raters. Results should also be separated by class and target size because a good aggregate score can hide overconfidence on small lesions. At present, the reviewed studies do not show that presenting several masks or an uncertainty heatmap improves clinical decisions, reading time, or patient outcomes.

4. Open challenges and future directions

Computation remains the most immediate deployment constraint. LDSeg and CRAC-DM show that latent and categorical formulations can shift the accuracy-speed trade-off [12,19]. Work outside segmentation suggests additional routes: Fast-DDPM aligned training and sampling at ten steps and reported about 100-fold faster sampling and five-fold faster training for medical image-to-image generation [28], while DeepCache reused temporally redundant high-level features and accelerated Stable Diffusion by 2.3 times with a small reported reduction in CLIP score [29]. These results are not segmentation benchmarks. Any transfer of such acceleration techniques must still demonstrate that small boundaries, class relations, sample diversity, and calibration are preserved.

Volumetric processing creates a second bottleneck. Independent 2D slices can disagree across the volume, while full 3D diffusion may exceed realistic memory limits. Latent states, categorical

DOI: 10.54254/2755-2721/2026.37692

labels, and region-focused refinement offer possible compromises. Papers should state clearly whether a method is genuinely volumetric or only combines 2D outputs. For deployment, performance on new scanners, centers, protocols, and populations is also more informative than another small improvement on a familiar public split.

Reproducibility is difficult because the final mask depends on more than the trained weights. Random seeds, noise schedules, sampling paths, thresholds, sample count, ensemble rules, and hardware all influence the output. Reporting should match the claimed role: direct generators need step count, sample count, latency, memory, and external results; representation studies need matched annotation budgets and backbones; refiners need before-and-after results for the same segmentor; and uncertainty studies need calibration and rater-distribution analysis. Exact inference settings, repeated-run variability, code, and trained weights where permitted would make the evidence easier to reproduce.

Anatomical plausibility is another weak point. Dice can stay high even when topology is broken, a structure has an impossible relation to a neighboring organ, or a thin connection disappears. Structural information could enter through conditioning, the latent space, an energy term, or a sampling rule. The aim should not be to suppress all diversity, but to keep alternative masks within anatomically defensible limits.

Clinical integration will likely require a division of labor between model families. The 2026 reviews continue to highlight efficiency, inconsistent reporting, explainability, and limited real-world validation [6-8]. Medical SAM 2, although not a diffusion model, treats 2D and 3D segmentation as video-style tracking with a self-sorting memory bank [30]. One practical hybrid would use a promptable foundation model for a rapid initial mask and reserve diffusion for difficult boundaries or ambiguous regions. Such systems should be compared with simpler prompt refinement and evaluated for interaction time, failure recovery, and clinician workload.

Human-centered evaluation should also move beyond visual preference. More useful outcomes include contour-editing time, referral for a second review, missed-lesion rates, inter-reader agreement, and unsafe automation bias. Interfaces must distinguish a set of plausible annotations from a calibrated probability of error. Deployment studies should record latency variation, available hardware, privacy constraints, and whether an output can be reproduced later for audit or regulation.

A role-based view provides a simple way to sharpen future studies. Authors can state whether diffusion is intended to generate, represent, refine, or quantify ambiguity and then choose the baseline and primary outcome accordingly. This makes it easier to identify a genuine role-specific benefit and harder to attribute every improvement to diffusion when a simpler component could have produced the same effect.

5. Limitations

This article is a focused narrative synthesis rather than an exhaustive inventory. Differences in datasets, preprocessing, target definitions, data splits, sampling schedules, ensemble sizes, and hardware prevent quantitative pooling or a universal ranking. SegDiff and Medical SAM 2 were still preprints in the versions examined, and evidence for representation reuse in medical segmentation remains partly indirect. The review also emphasizes representative methods and reported findings rather than conducting a formal risk-of-bias assessment. Finally, most primary studies rely on retrospective public data, so prospective evidence of clinical benefit is still limited.

DOI: 10.54254/2755-2721/2026.37692

6. Conclusion

Diffusion is best viewed as a set of roles rather than a single segmentation strategy. The strongest evidence supports direct generation and prediction refinement, while representation reuse is less directly validated and uncertainty claims require calibration or agreement with expert variation. The practical implication is role-matched evaluation: generators must justify sampling cost, refiners must improve the same base model, and uncertainty methods must show that variation is meaningful. Efficient 3D inference, external validation, anatomical validity, and prospective human evaluation will determine whether these methods become useful clinical components rather than computationally expensive alternatives.

References

[1] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015*. Munich, Germany. pp. 234-241. https://doi.org/10.1007/978-3-319-24574-4_28.

[2] Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems 33*. Online. pp. 6840-6851.

[3] Yang, L., Zhang, Z., Song, Y., et al. (2023). Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys, 56*(4): 1-39. https://doi.org/10.1145/3626235.

[4] Kazerouni, A., Aghdam, E. K., Heidari, M., et al. (2023). Diffusion models in medical imaging: A comprehensive survey. *Medical Image Analysis, 88*: 102846. https://doi.org/10.1016/j.media.2023.102846.

[5] Shi, Y., Abulizi, A., Wang, H., et al. (2025). Diffusion models for medical image computing: A survey. *Tsinghua Science and Technology, 30*(1): 357-383. https://doi.org/10.26599/TST.2024.9010047.

[6] Yaseen, M., Ali, M., Ali, S., et al. (2026). Diffusion-based approaches for medical image segmentation: An in-depth review. *Electronics, 15*(7): 1400. https://doi.org/10.3390/electronics15071400.

[7] Azad, M., Fahad, N. M., Raiaan, M. A. K., et al. (2026). A systematic review of diffusion models for medical image-based diagnosis: Methods, taxonomies, clinical integration, explainability, and future directions. *Diagnostics, 16*(2), 211. https://doi.org/10.3390/diagnostics16020211.

[8] Sabanci, K., Aslan, B., & Aslan, M. F. (2026). Medical image segmentation methods: A decision-guided survey covering 2D/3D CNNs, transformers, VLMs, SAM-based models and diffusion approaches. *Bioengineering, 13*(5), 555. https://doi.org/10.3390/bioengineering13050555.

[9] Nichol, A. Q., & Dhariwal, P. (2021). Improved denoising diffusion probabilistic models. In *Proceedings of the 38th International Conference on Machine Learning* (pp. 8162–8171). Online.

[10] Rombach, R., Blattmann, A., Lorenz, D., et al. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 10684–10695). New Orleans, USA. https://doi.org/10.1109/CVPR52688.2022.01042.

[11] Chen, T., Wang, C., Chen, Z., et al. (2024). HiDiff: Hybrid diffusion framework for medical image segmentation. *IEEE Transactions on Medical Imaging, 43*(10), 3570–3583. https://doi.org/10.1109/TMI.2024.3424471.

[12] Zhou, Y., Xu, C., Awad, Z., et al. (2026). CRAC-DM: Class relation-aware categorical diffusion model for surgical scene segmentation. *International Journal of Computer Assisted Radiology and Surgery, 21*, 949–958. https://doi.org/10.1007/s11548-026-03601-7.

[13] Choi, J., Lee, J., Shin, C., et al. (2022). Perception prioritized training of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. New Orleans, USA. pp. 11472-11481.

[14] Amit, T., Shaharbany, T., Nachmani, E., et al. (2022). SegDiff: Image segmentation with diffusion probabilistic models. https://arxiv.org/abs/2112.00390.

[15] Wu, J., Fu, R., Fang, H., et al. (2024). MedSegDiff: Medical image segmentation with diffusion probabilistic model. In *Medical Imaging with Deep Learning*. Nashville, USA. pp. 1623-1639.

[16] Wu, J., Ji, W., Fu, H., et al. (2024). MedSegDiff-V2: Diffusion-based medical image segmentation with transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*. Vancouver, Canada. pp. 6030-6038. https://doi.org/10.1609/aaai.v38i6.28418.

[17] Xing, Z., Wan, L., Fu, H., et al. (2025). Diff-UNet: A diffusion embedded network for robust 3D medical image segmentation. *Medical Image Analysis, 105*: 103654. https://doi.org/10.1016/j.media.2025.103654.

DOI: 10.54254/2755-2721/2026.37692

[18] Wu, X., Hou, Q., Xu, Z., et al. (2025). FCFDiff-Net: Full-conditional feature diffusion embedded network for 3D brain tumor segmentation. *Quantitative Imaging in Medicine and Surgery, 15*(5): 4217-4234. https://doi.org/10.21037/qims-24-2300.

[19] Zaman, F. A., Jacob, M., Chang, A., et al. (2026). Latent diffusion for medical image segmentation: End-to-end learning for fast sampling and accuracy. *Biomedical Signal Processing and Control, 114*: 109380. https://doi.org/10.1016/j.bspc.2025.109380.

[20] Guo, X., Yang, Y., Ye, C., et al. (2023). Accelerating diffusion models via pre-segmentation diffusion sampling for medical image segmentation. In: *2023 IEEE 20th International Symposium on Biomedical Imaging*. Cartagena de Indias, Colombia. pp. 1-5. https://doi.org/10.1109/ISBI53787.2023.10230524.

[21] Liu, X., Li, W., Yuan, Y. (2024). DiffRect: Latent diffusion label rectification for semi-supervised medical image segmentation. In: *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2024*. Marrakesh, Morocco. pp. 56-66. https://doi.org/10.1007/978-3-031-72390-2_6.

[22] Rahman, A., Valanarasu, J. M. J., Hacihaliloglu, I., et al. (2023). Ambiguous medical image segmentation using diffusion models. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Vancouver, Canada. pp. 11536-11546.

[23] Tian, J., Aggarwal, L., Colaco, A., et al. (2024). Diffuse, attend, and segment: Unsupervised zero-shot segmentation using Stable Diffusion. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Seattle, USA. pp. 3554-3563.

[24] Graikos, A., Malkin, N., Jojic, N., et al. (2022). Diffusion models as plug-and-play priors. In: *Advances in Neural Information Processing Systems 35*. New Orleans, USA. pp. 14715-14728.

[25] Khader, F., Müller-Franzes, G., Tayebi Arasteh, S., et al. (2023). Denoising diffusion probabilistic models for 3D medical image generation. *Scientific Reports, 13*: 7303. https://doi.org/10.1038/s41598-023-34341-2.

[26] Dorjsembe, Z., Pao, H.-K., Odonchimed, S., et al. (2024). Conditional diffusion models for semantic 3D brain MRI synthesis. *IEEE Journal of Biomedical and Health Informatics, 28*(7): 4084-4093. https://doi.org/10.1109/JBHI.2024.3385504.

[27] Huang, L., Ruan, S., Decazes, P., et al. (2025). Deep evidential fusion with uncertainty quantification and reliability learning for multimodal medical image segmentation. *Information Fusion, 113*: 102648. https://doi.org/10.1016/j.inffus.2024.102648.

[28] Jiang, H., Imran, M., Ma, L., et al. (2025). Fast-DDPM: Fast denoising diffusion probabilistic models for medical image-to-image generation. *IEEE Journal of Biomedical and Health Informatics, 29*(10): 7326-7335. https://doi.org/10.1109/JBHI.2025.3565183.

[29] Ma, X., Fang, G., Wang, X. (2024). DeepCache: Accelerating diffusion models for free. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Seattle, USA. pp. 15762-15772.

[30] Zhu, J., Hamdi, A., Qi, Y., et al. (2024). Medical SAM 2: Segment medical images as video via Segment Anything Model 2. https://arxiv.org/abs/2408.00874.