Shape-Scale Aware Diffusion Models for Weakly Supervised Medical Image Anomaly Detection

Shenhai Zheng[1,2], Dongtao Tan[1], Weisheng Li[1,2], and Laquan Li[2,3](✉)

[1] School of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, China
[2] Chongqing College of Artificial Intelligence, Chongqing 401339, China
[3] School of Mathematics and Statistics, Chongqing University of Posts and Telecommunications, Chongqing 400065, China
lilq@cqupt.edu.cn

Abstract. Recently, diffusion models have shown promising performance in medical image anomaly detection. However, most existing methods insufficiently model complex anatomical structures and irregular lesion morphologies. To address this limitation, we propose a *Direction-Guided Shape-Scale Aware Diffusion Model* for medical anomaly detection. The proposed model introduces a *Direction-Guided Multi-Scale* residual module to decouple directional features and aggregate multi-scale receptive fields, thereby enhancing anatomical shape representation. In addition, an *Anomaly-Sensitive Skip Fusion* strategy is developed to guide feature fusion using high-frequency structural priors, enabling the network to better focus on fine-grained lesion boundaries and abnormal details. Furthermore, a lightweight architecture is further adopted to reduce the computational burden of iterative diffusion inference. Experimental results demonstrate that the proposed method achieves superior detection accuracy and inference efficiency compared with existing approaches. Our code is available at https://github.com/guyxlouspg/AnoSSPD.

Keywords: Anomaly detection · Medical images · Diffusion models · Shape-scale aware.

1 Introduction

By deducing pixel-level localizations from cost-effective image-level labels, weakly supervised anomaly detection strikes an optimal balance between prohibitive fully supervised costs and underperforming unsupervised methods [15,16,19,20]. However, current diffusion-based approaches overlook the structural flaws of standard U-Net backbones: (1) isotropic convolutions fail to capture complex lesion morphologies; (2) indiscriminate skip connections leak anomalous features into healthy reconstructions; and (3) heavy parameter counts exacerbate the latency of iterative denoising.

Early medical anomaly detection relied on VAEs [11, 21–23] and GANs [7, 17,18] to model healthy data distributions. However, VAEs often produce overly

smooth reconstructions lacking high-frequency details, while GANs suffer from mode collapse. Consequently, Denoising Diffusion Probabilistic Models (DDPMs) have emerged as the dominant paradigm due to their superior synthesis fidelity.

Recent diffusion-based anomaly detection primarily optimizes the diffusion process itself. Unsupervised methods enhance computational efficiency and anomaly capture [2–5, 9, 15, 20], while weakly supervised approaches leverage image-level labels to guide the diffusion trajectory toward healthy domains [6, 8, 13, 16, 19]. Despite these algorithmic advancements in sampling and guidance mechanisms, most methods predominantly rely on standard U-Net backbones. Consequently, they overlook the architectural adaptability required to model complex, multi-scale, and directionally variant anatomical structures—a critical gap our framework directly resolves.

To resolve these architectural limitations, we propose a Morphology-Scale Aware Lightweight Backbone for medical anomaly detection. First, the Direction-guided Multi-Scale (DMS) module utilizes dual-branch dilated convolutions and hierarchical SE attention to capture direction-decoupled features and global anatomical constraints. Second, the Anomaly-Sensitive Skip Fusion (ASSF) mechanism dynamically gates skip connections using high-frequency boundary descriptors to prevent anomaly leakage. Finally, the entire network is constructed upon Point-Depth Shuffle Convolutions (PDS-Conv), decoupling spatial and channel perception to drastically reduce computational overhead while preserving essential long-range contexts. Our main contributions are threefold: (1) We propose the Direction-guided Multi-Scale (DMS) residual module. Through the synergy of multi-scale fields of view and direction-decoupled attention, it establishes robust shape-scale perception capabilities. (2) We design an Anomaly-Sensitive Skip Fusion (ASSF) mechanism. By utilizing high-frequency boundary cues to adaptively gate skip connection features, it significantly improves detection sensitivity regarding lesion boundaries. (3) We construct a lightweight diffusion backbone based on PDS-Conv. This architecture retains the essential long-range anatomical context for lesion identification while achieving a substantial reduction in both model parameters and computational complexity.

2 Methods

2.1 Overall Architecture

Fig. 1 illustrates our AnoSSPD. The architecture integrates three core components: (1) LightResBlocks deployed across the encoder and decoder stages to minimize computational redundancy while preserving feature extraction; (2) a Direction-guided Multi-scale (DMS) ResBlock at the bottleneck to capture global shape-scale contexts and establish robust anatomical constraints; and (3) Anomaly-Sensitive Skip Fusion (ASSF) mechanisms replacing standard skip connections, which boost anomaly sensitivity via high-frequency gating. Finally, during inference, we employ AnoFPDM's forward noising process [6] to generate and aggregate single-step anomaly sub-maps into the final prediction.

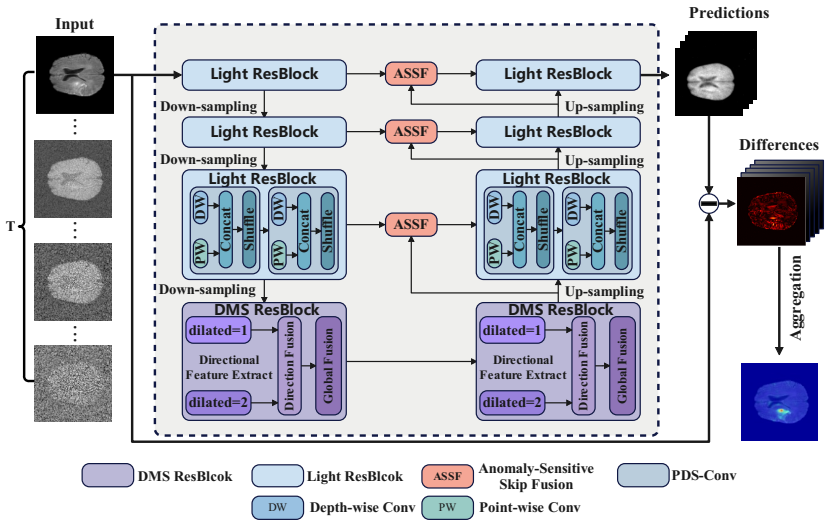

Fig. 1: The overall framework of the proposed Anomaly Shape-Scale Perception Diffusion Model (AnoSSPD).

2.2 Direction-guided Multi-scale (DMS) Residual Module

To overcome the limited contextual modeling of standard deep networks and the loss of fine anatomical structures during spatial down-sampling, we propose the Direction-Guided Multi-Scale (DMS) residual module. To expand the receptive field without resolution degradation, we employ a dual-branch parallel dilated convolution structure (dilation rates $d_k \in \{1, 2\}$) . However, standard dilated convolutions suffer from the "grid effect", which disrupts local pixel correlations—a critical flaw given the topological continuity of medical anatomy. To mitigate this and capture structural orientations, we embed independent direction-guided convolutions within each dilation branch. Specifically, feature extraction is decoupled into horizontal, vertical, and diagonal orientations ($R = \{\text{hor}, \text{ver}, \text{diag}\}$):

$$h_r^{d_k} = \text{DC}_r^{d_k}(h), \quad r \in R, \quad k \in 1, 2, \tag{1}$$

where h is the input feature and DC denotes the directional convolution.

To effectively integrate these heterogeneous, multi-scale, and directionally decoupled features, we design a Hierarchical Adaptive Feature Fusion mechanism. Features sharing the same directional attribute r across different scales d_k are first concatenated and refined using Direction-Level Squeeze-and-Excitation (SE_R) attention. Subsequently, all direction-specific features are concatenated and globally recalibrated via Global SE (SE_G) attention to establish a holistic

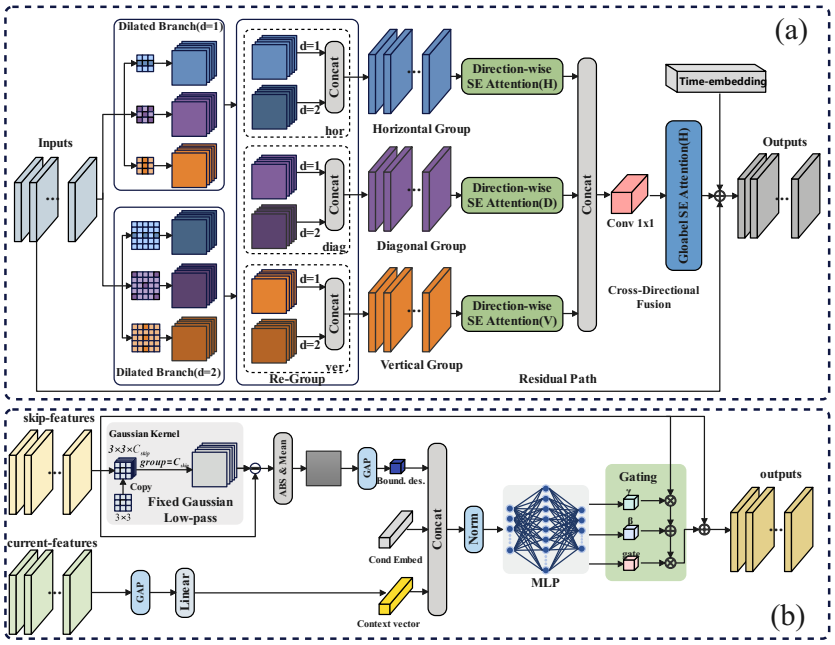

Fig. 2: (a)Schematic diagram of the direction-guided multi-scale residual module (DMS). (b)Illustration of the Anomaly-Sensitive Skip Fusion (ASSF) module.

structural representation h':

$$\begin{aligned} h_r &= \mathrm{Concat}(h_r^{d_1}, h_r^{d_2}), \quad \tilde{h}_r = \mathrm{SE}_R(h_r), \\ h' &= \mathrm{SE}_G(\mathrm{Fusion}(\mathrm{Concat}(\tilde{h}_{\mathrm{hor}}, \tilde{h}_{\mathrm{ver}}, \tilde{h}_{\mathrm{diag}}))). \end{aligned} \tag{2}$$

This hierarchical process mathematically enriches multi-scale details and preserves topological continuity, providing robust structural guidance for the diffusion reconstruction. Fig. 2(a) illustrates this mechanism in detail.

2.3 Anomaly-Sensitive Skip Fusion

Standard U-Net skip connections indiscriminately concatenate features, diluting high-frequency pathological details with background noise. To resolve this, our Anomaly-Sensitive Skip Fusion (ASSF) dynamically gates encoder features based on the decoder's generative state.

First, we extract a context vector c_{ctx} via Global Average Pooling (GAP) on the current decoder feature h_{dec}, which captures the ongoing reconstruction state. c_{ctx} then acts as a top-down semantic condition to inform the gating mechanism of current structural needs. Concurrently, to explicitly isolate anomalous

high-frequency cues, we subtract a predefined Gaussian low-pass background from the normalized skip feature h_{skip} to derive a structural descriptor e_{des}:

$$e_{\text{des}} = \text{GAP}\left(\text{mean}\left(|h_{\text{skip}} - W_{\text{Gauss}} * h_{\text{skip}}|\right)\right). \tag{3}$$

A unified condition vector c_{cond}—formed by concatenating the timestep embedding c_{emb}, c_{ctx}, and e_{des}—is then mapped via an MLP to channel-wise modulation parameters $[\alpha, \gamma, \beta]$:

$$[\alpha, \gamma, \beta] = \text{MLP}(\text{LayerNorm}(c_{\text{cond}})). \tag{4}$$

Finally, h_{skip} undergoes an affine transformation (γ, β), selective gating (α), and a residual connection to yield the noise-filtered, refined feature $\hat{h}_{\text{skip}}$:

$$\hat{h}_{\text{skip}} = h_{\text{skip}} + \alpha \odot (\gamma \odot h_{\text{skip}} + \beta). \tag{5}$$

By filtering out irrelevant background perturbations, ASSF ensures that only clean, structurally significant features are propagated to the decoder. Fig. 2(b) illustrates this mechanism in detail.

2.4 Lightweight Network Architecture

To mitigate the prohibitive computational overhead of iterative diffusion, we introduce the Point-Depth Shuffle Convolution (PDS-Conv), inspired by [12]. PDS-Conv first compresses input channels by half via a point-wise (PW) convolution for efficient inter-channel exchange. Concurrently, an auxiliary 5×5 depth-wise (DW) branch captures long-range anatomical dependencies without parameter inflation. To resolve the depth-wise "information gap", PW and DW outputs are concatenated and channel-shuffled, seamlessly fusing dense channel correlations with rich spatial features. By replacing the expensive stacked 3×3 convolutions in standard U-Net ResBlocks with PDS-Conv, we construct the LightResBlock. This design fully preserves residual and timestep embedding mechanisms, drastically reducing FLOPs and parameters while maintaining high-fidelity feature extraction.

3 Experiments

3.1 Dataset and Setup

We evaluated our method on three public datasets: BraTS 2021 [1], BraTS-PEDs 2024 [10], and ATLAS v2.0 [14]. **BraTS 2021**: Comprises 1,254 multi-modal (T1, T1ce, T2, FLAIR) 3D MRI scans from glioma patients. Scans were processed into 2D axial slices. Following [19], we discarded background-only and extreme slices (bottom 80, top 26). Valid slices were resized to 128 x 128, normalized to a 0-1 range, and binarized for general anomaly detection (healthy vs. unhealthy). While all four modalities are stacked during training, only FLAIR

and T2 are utilized for inference. **BraTS-PEDs 2024**: Contains 457 pediatric cases presenting distinct spatial and texture patterns. Preprocessing is identical to BraTS 2021, except no slice exclusion was applied. **ATLAS v2.0**: Includes 1,271 T1-weighted scans for stroke lesion identification. Preprocessing aligns with BraTS 2021, but experiments are strictly conducted in a single-modality setting.

Adhering to the AnoFPDM forward noising framework [6], datasets were partitioned into training, validation, and testing sets at a subject-level ratio of 75:5:20. Models were trained on a single NVIDIA Tesla T4 GPU for 140,000 iterations with a batch size of 16. For evaluation efficiency, 200 samples each from the validation and test sets were randomly selected for hyperparameter tuning and final evaluation, respectively.

3.2 Main Results

We evaluated our method against unsupervised baselines—specifically AnoDDPM utilizing standard Gaussian noise AnoDDPM(G) and Simplex noise AnoDDPM(s) as well as weakly supervised methods (Clf-Guided, Diff-SCM, AnoFPDM) using pixel-level Dice, IoU, and AUPRC under "Mixed" (overall capability) and "Unhealthy" (lesion sensitivity) splits.

We employ pixel-level metrics to evaluate model performance, including DICE score and Area Under the Precision-Recall Curve (AUPRC), with a specific focus on foreground region. We evaluated these metrics under two data settings: Mixed (all samples) and Unhealthy (unhealthy samples only). The former measures overall detection capability, while the latter rigorously assesses the model's sensitivity and segmentation accuracy in pathological regions. Table 1 summarized the quantitative results. In the Unhealthy configuration, our model demonstrates even more significant improvements, indicating that our shape-scale-aware architecture substantially enhances precise localization of lesion boundaries and exhibits more robust performance even on highly challenging datasets.

Furthermore, visually results (Fig. 3) show that while baselines suffer from fragmented boundaries or under-segmented complex tumor cores, our approach precisely delineates complete lesion structures and effectively suppresses false positives in healthy tissues. AnoSSPD exhibits exceptional robustness against severe data challenges.

Table 1: Quantitative comparison results. The best results are highlighted in **bold**, and the second-best results are underlined.

Methods	**BraTS 2021**				**BraTS-PEDs 2024**				**ATLAS v2.0**			
	Mixed		Unhealthy		Mixed		Unhealthy		Mixed		Unhealthy	
	DICE	AUPRC	DICE	AUPRC	DICE	AUPRC	DICE	AUPRC	DICE	AUPRC	DICE	AUPRC
AnoDDPM(G) [20]	49.0	32.6	25.6	44.3	62.2	6.3	8.0	16.3	41.2	4.8	7.5	10.6
AnoDDPM(S) [20]	67.9	70.5	50.3	75.2	61.6	28.2	28.1	43.5	57.3	33.9	18.3	41.8
Diff-SCM [16]	74.3	69.7	61.5	78.3	66.3	24.1	22.4	32.3	16.0	1.4	2.4	3.2
Clf-Guided [19]	71.1	79.1	63.9	82.8	**74.3**	17.4	19.5	24.7	46.5	3.7	2.2	7.7
AnoFPDM [6]	75.3	75.6	64.9	81.6	72.6	33.0	43.2	45.3	58.0	27.8	20.1	30.0
Ours	**77.2**	**79.4**	**68.7**	**84.9**	67.7	**63.7**	**48.4**	**71.1**	**58.4**	**45.7**	**26.7**	**47.6**

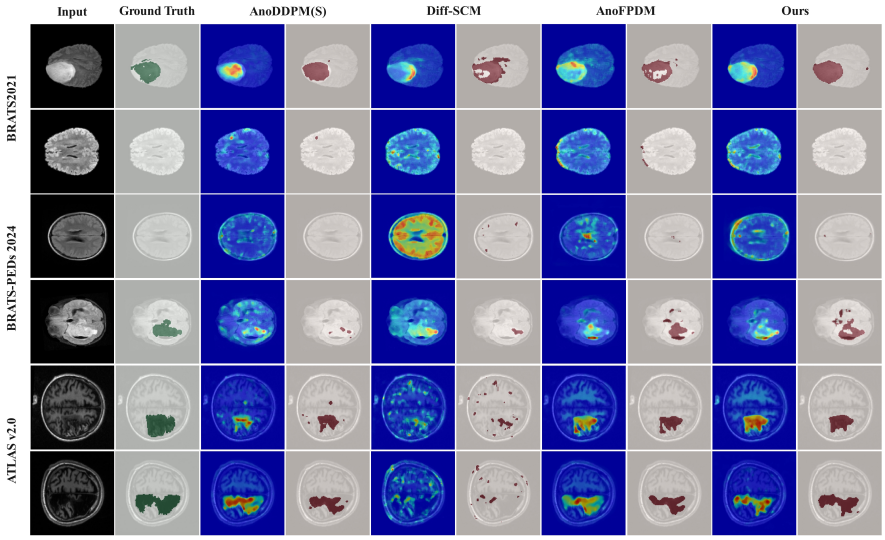

Fig. 3: Visualization results of anomaly detection across three datasets: BraTS2021, BraTS-PEDs 2024, and ATLAS v2.0. Red regions in the heatmaps indicate areas predicted as anomalies by the models.

3.3 Ablation Study

To validate our architectural contributions, we conducted ablation studies on the BraTS 2021 dataset. As shown in Table 2, deploying the DMS module exclusively at the bottleneck layer yields the optimal performance-efficiency trade-off.

Extending it to high-resolution decoder layers (Extensions 1–3) degrades several key indicators, implying that while DMS effectively captures high-level semantics, its application to lower-level features interferes with local detail preservation. Furthermore, replacing standard skip connections with our ASSF mechanism effectively suppresses irrelevant background noise while highlighting crucial pathological structures, proving its indispensability.

Table 3 evaluates our lightweight PDS-Conv design. Counterintuitively, deploying PDS-Conv at both the input and output layers of the ResBlock (PDSConv (in+out)) outperforms its partial application (PDSConv(in)) despite being lighter. This suggests that mixing standard convolutions with PDS-Conv disrupts feature consistency, whereas a fully unified block maintains high-fidelity extraction.

As detailed in Table 4, our lightweight architecture drastically reduces model parameters by 52.9% and inference computational overhead (GFLOPs) by 77.3% compared to the baseline, without compromising detection performance. This substantial efficiency gain significantly surpasses existing baselines, validating the robustness of our module design and facilitating highly viable, low-cost deployment in resource-constrained clinical settings.

Table 2: Ablation study on BraTS2021. "Bottleneck"/"Layer 1–3": DMS module applied at the bottleneck / extended to successive encoder and decoder layers.

Components					Mixed			Unhealthy		
Bottleneck	Layer1	Layer2	Layer3	ASSF	DICE	IoU	AUPRC	DICE	IoU	AUPRC
					75.3	69.7	75.6	64.9	55.1	81.6
✓					76.2	70.3	75.4	66.9	56.7	82.3
✓	✓				<u>77.1</u>	70.8	74.8	65.8	55.5	80.9
✓	✓	✓			73.4	67.5	<u>78.4</u>	<u>68.4</u>	<u>57.9</u>	<u>83.6</u>
✓	✓	✓	✓		77.0	<u>71.1</u>	74.8	66.3	56.3	83.0
✓				✓	**77.2**	**71.1**	**79.4**	**68.7**	**59.1**	**84.9**

Table 3: Performance evaluation of different lightweight architecture strategies. "PDSConv (in)" denotes applying PDSConv only at the input projection of the residual block, while "PDSConv (in+out)" indicates its application at both input and output projections.

Methods	Mixed			Unhealthy		
	DICE	IoU	AUPRC	DICE	IoU	AUPRC
Baseline	**77.2**	<u>71.1</u>	**79.4**	**68.7**	**59.1**	**84.9**
PDSConv (in)	76.0	70.0	75.3	67.1	57.2	82.5
PDSConv (in+out)	<u>77.1</u>	**71.6**	<u>78.6</u>	<u>68.2</u>	<u>57.6</u>	<u>83.7</u>

Table 4: Comparison of model complexity and computational cost.

Method	Params (M)	GFLOPs/img
AnoDDPM [20]	91.85	65.03
CLf-Guided [19]	91.85	64.94
Diff-SCM [16]	98.30	64.92
AnoFPDM [6]	98.30	64.92
Our Baseline	97.89	65.00
PDSConv (in)	<u>66.62</u>	<u>33.99</u>
PDSConv (in+out)	**46.12**	**14.75**

4 Conclusion

We propose AnoSSPD to address the challenges of complex lesion morphologies and high computational costs in weakly supervised medical anomaly detection. By integrating the DMS module, ASSF mechanism, and a lightweight PDS-Conv backbone, AnoSSPD achieves robust, SOTA localization while significantly reducing computational overhead by 77.3%. While currently limited to a binary segmentation paradigm that overlooks semantic distinctions between pathological subtypes, future work will extend the framework to multi-class anomaly detection, enabling refined pathological subtyping for precise clinical diagnosis.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China(Nos. 62331008, 62221005, 61902046 and 61901074), the Science and Technology Research Program of Chongqing Municipal Education Commission (No. KJZD-M202620605) and the Natural Science Foundation of Chongqing (No. 2023NSCQ-LZX0045).

Disclosure of Interests

The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Baid, U., Ghodasara, S., Mohan, S., Bilello, M., Calabrese, E., Colak, E., Farahani, K., Kalpathy-Cramer, J., Kitamura, F.C., Pati, S., et al.: The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. arXiv preprint arXiv:2107.02314 (2021)
2. Baugh, M., Reynaud, H., Marimont, S.N., Cechnicka, S., Müller, J.P., Tarroni, G., Kainz, B.: Image-conditioned diffusion models for medical anomaly detection. In: International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging. pp. 117–127. Springer (2024)
3. Behrendt, F., Bhattacharya, D., Krüger, J., Opfer, R., Schlaefer, A.: Patched diffusion models for unsupervised anomaly detection in brain mri. In: Medical imaging with deep learning. pp. 1019–1032. PMLR (2024)
4. Bercea, C.I., Neumayr, M., Rueckert, D., Schnabel, J.A.: Mask, stitch, and re-sample: Enhancing robustness and generalizability in anomaly detection through automatic diffusion models (2023)
5. Bercea, C.I., Wiestler, B., Rueckert, D., Schnabel, J.A.: Diffusion models with implicit guidance for medical anomaly detection. In: Medical Image Computing and Computer Assisted Intervention. pp. 211–220 (2024)
6. Che, Y., Rafsani, F., Shah, J., Siddiquee, M.M.R., Wu, T.: Anofpdm: Anomaly detection with forward process of diffusion models for brain mri. In: Proceedings of the Winter Conference on Applications of Computer Vision (WACV) Workshops. pp. 1113–1122 (February 2025)
7. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in Neural Information Processing Systems. vol. 27 (2014)
8. Hu, X., Chen, Y.J., Ho, T.Y., Shi, Y.: Conditional diffusion models for weakly supervised medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention. pp. 756–765 (2023)
9. Iqbal, H., Khalid, U., Chen, C., Hua, J.: Unsupervised anomaly detection in medical images using masked diffusion model. In: Cao, X., Xu, X., Rekik, I., Cui, Z., Ouyang, X. (eds.) Machine Learning in Medical Imaging. pp. 372–381. Springer Nature Switzerland, Cham (2024)

10. Kazerooni, A.F., Khalili, N., Liu, X., Gandhi, D., Jiang, Z., Anwar, S.M., Albrecht, J., Adewole, M., Anazodo, U., Anderson, H., et al.: The brain tumor segmentation in pediatrics (BraTS-PEDs) challenge: focus on pediatrics (CBTN-CONNECT-DIPGR-ASNR-MICCAI BraTS-PEDs). arXiv preprint arXiv:2404.15009 (2024)
11. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
12. Li, H., Li, J., Wei, H., Liu, Z., Zhan, Z., Ren, Q.: Slim-neck by gsconv: A lightweight-design for real-time detector architectures. Journal of Real-Time Image Processing **21**(3), 62 (2024)
13. Li, J., Cao, H., Wang, J., Liu, F., Dou, Q., Chen, G., Heng, P.A.: Fast non-markovian diffusion model for weakly supervised anomaly detection in brain mr images. In: Medical Image Computing and Computer Assisted Intervention. pp. 579–589 (2023)
14. Liew, S.L., Lo, B.P., Donnelly, M.R., Zavaliangos-Petropulu, A., Jeong, J.N., Barisano, G., Hutton, A., Simon, J.P., Juliano, J.M., Suri, A., et al.: A large, curated, open-source stroke neuroimaging dataset to improve lesion segmentation algorithms. Scientific data **9**(1), 320 (2022)
15. Pinaya, W.H.L., Graham, M.S., Gray, R., da Costa, P.F., Tudosiu, P.D., Wright, P., Mah, Y.H., MacKinnon, A.D., Teo, J.T., Jager, R., Werring, D., Rees, G., Nachev, P., Ourselin, S., Cardoso, M.J.: Fast unsupervised brain anomaly detection and segmentation with diffusion models. In: Medical Image Computing and Computer Assisted Intervention. pp. 705–714 (2022)
16. Sanchez, P., Kascenas, A., Liu, X., O'Neil, A.Q., Tsaftaris, S.A.: What is healthy? generative counterfactual diffusion for lesion localization. In: Mukhopadhyay, A., Oksuz, I., Engelhardt, S., Zhu, D., Yuan, Y. (eds.) Deep Generative Models. pp. 34–44. Springer Nature Switzerland, Cham (2022)
17. Schlegl, T., Seeböck, P., Waldstein, S.M., Langs, G., Schmidt-Erfurth, U.: f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks. Medical Image Analysis **54**, 30–44 (2019)
18. Siddiquee, M.M.R., Zhou, Z., Tajbakhsh, N., Feng, R., Gotway, M.B., Bengio, Y., Liang, J.: Learning fixed points in generative adversarial networks: From image-to-image translation to disease detection and localization. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 191–200 (2019)
19. Wolleb, J., Bieder, F., Sandkühler, R., Cattin, P.C.: Diffusion models for medical anomaly detection. In: Medical Image Computing and Computer Assisted Intervention. pp. 35–45 (2022)
20. Wyatt, J., Leach, A., Schmon, S.M., Willcocks, C.G.: Anoddpm: Anomaly detection with denoising diffusion probabilistic models using simplex noise. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 650–656 (2022)
21. Zhou, L., Deng, W., Wu, X.: Unsupervised anomaly localization using vae and beta-vae. arXiv preprint arXiv:2005.10686 (2020)
22. Zimmerer, D., Isensee, F., Petersen, J., Kohl, S., Maier-Hein, K.: Unsupervised anomaly localization using variational auto-encoders. In: Medical Image Computing and Computer Assisted Intervention. pp. 289–297 (2019)
23. Zimmerer, D., Kohl, S., Petersen, J., Isensee, F., Maier-Hein, K.: Context-encoding variational autoencoder for unsupervised anomaly detection–short paper. arXiv preprint arXiv:1907.12258 (2019)