Noisy Label Refinement Based on Discrete Diffusion Process in 3D Ossicle Segmentation

Linqian Fan[1⋆], Mengshi Zhang[2⋆], Yonghao Wang[1], Wenkai Lu[1✉], and Hongxia Yin[1,2,3✉]

[1] Department of Automation, Tsinghua University, Beijing, China
[2] Department of Radiology, Beijing Friendship Hospital, Capital Medical University, Beijing, China
[3] Department of Medical Engineering, Beijing Friendship Hospital, Capital Medical University, Beijing, China
lwkmf@mail.tsinghua.edu.cn, 282496774@qq.com

Abstract. Ossicular chain lesions can cause hearing loss, making accurate segmentation of ossicles critical for clinical diagnosis and treatment. Ultra-high-resolution computed tomography (U-HRCT) provides quality images for ossicle segmentation tasks, but the complex structure of the stapes and variations in annotators' experience often lead to noisy labels in 3D annotation within clinical practice. To address this, we propose a novel framework tailored for two types of noisy labels: (1) *incomplete-structure* labels, and (2) *complete-structure* but *inaccurate* labels. For the former, we introduce a Dilating&Selecting (D&S) framework, which completes missing structures using a dilating Volumetric Discrete Diffusion Refiner (VDDR) with a novel cover loss and evaluates label completeness via a completeness selection strategy. For the latter, we introduce a noise-based augmentation to better train VDDR. Experimental results demonstrate that D&S framework reduce the time cost of manual annotation by 90.2%, while VDDR outperforms other state-of-the-art methods. To facilitate further research and development, our code and two datasets are publicly available.

Keywords: ossicles segmentation · noisy label · discrete diffusion model.

1 Introduction

The ossicular chain is a small, complex structure in the middle ear, comprising three bones: the malleus, incus, and stapes [14]. The stapes, the smallest bone in the body, have a base plate about 0.3 mm thick [6]. Once the ossicular chain is damaged, it may lead to sound conduction disorders and varying degrees of hearing loss, seriously affecting the physical and mental health of patients [19]. The segmentation of the ossicular chain is of great significance for medical imaging diagnosis and surgical planning [21]. Previous studies [15,8,16] have focused on the ossicular chain segmentaiton, however, conventional CT imaging struggles

⋆ These authors contributed equally to this work.

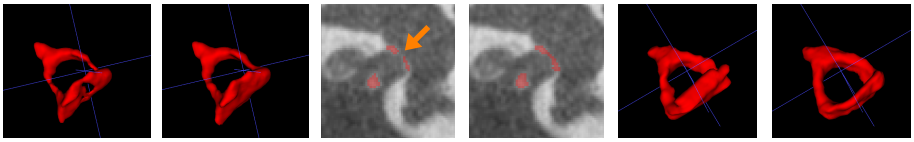

Fig. 1. Illustrations of different types of noisy labels compared with their corresponding ground truth (GT).

with accurate identification, particularly of the stapes. Ultra-high-resolution CT (U-HRCT) offers clearer imaging, improving ossicular chain visualization [25]. However, the complexity of the stapes and variability in annotator experience often result in inaccurate annotations, which are referred to as noisy labels [20]. In medical image computing and related fields, Ground Truth (GT), which refers to the actual labels, typically requires experts to consume substantial time for meticulous annotation. Consequently, obtaining authentic ossicle segmentation labels on large-scale datasets is often impractical. Thus, developing more precise automatic segmentation algorithms and tools is vital for improving ossicular chain annotation accuracy, efficiency, and enabling automated measurement and diagnosis.

Unlike previous works [18,24,7] that attribute noisy labels as ambiguous errors caused by human or machine annotators, this paper discusses noisy labels where the causes are known. The practical scenario we consider is as follows: Two datasets, **OSS-I** and **OSS-C**, need to be annotated. Due to the extremely small size of the stapes in 3D CT scans (only contains 1-2 pixels in some slices) and low signal-to-noise ratio (SNR), the annotators must decide whether to segment uncertain pixels based on 2D views. For **OSS-I**, annotators were instructed to mark only the clearly visible parts of the stapes, resulting in an *incomplete-structure* label compared to the GT, as shown in Figure 1(a) and (b). Figure 1(c) shows a representative slice from Figure 1(a), illustrating how noise at the base of the stapes (highlighted by the orange arrow) leads to missing annotations in that region. For **OSS-C**, annotators were instead asked to complete the stapes structure based anatomical knowledge, marking as much as possible. The resulting *complete-structure* annotation (Figure 1(e)) is more complete but still not fully accurate compared to the GT (Figure 1(f)). We define GT as the segmentation results that are deemed satisfactory by relevant expert annotators. It is relatively easy for experts to judge whether a result is acceptable, but it is much more difficult and time-consuming for them to manually segment satisfactory results. So we wanted to reduce the labeling burden by our methods. The noisy labels in OSS-I and OSS-C result from practical limitations, both differ from the expert-approved GT. Since generating GT is time-consuming, we train models to refine noisy labels, allowing experts to focus on verification rather than full annotation.

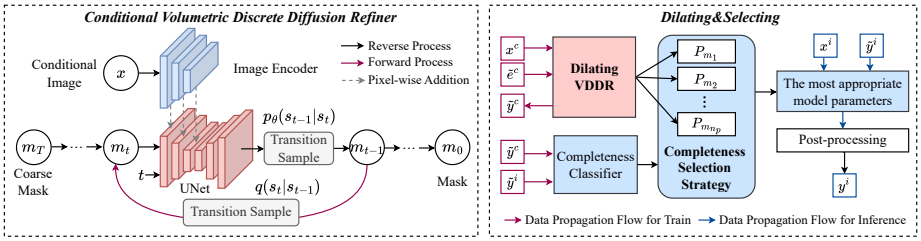

Fig. 2. Overview of the proposed conditional **Volumetric Descrete Diffusion Refiner (VDDR)** and **Dilating&Selecting (D&S)** framework.

To correct the aforementioned noisy labels, we draw inspiration from diffusion models [13,1] and its application on medical segmention workflow [3,9]. We conceptualize label refinement as a data generation process, where the diffusion model progressively denoises the noisy labels, thereby making the refinement process more stable and controllable. Motivated by [22], we proposed a conditional Volumetric Discrete Diffusion Refiner (VDDR) which is more suitable for segmentation due to the discrete diffusion process. Furthermore, VDDR is model-agnostic and is convenient to be applied in different situations. We first design a Dilating & Selecting framework which contains a dilating VDDR to expand and complete the *incomplete-structure* labels with a novel cover loss and a completeness selection strategy to select the appropriate model parameters. After obtaining partially accurate labels, we use VDDR to correct the inaccurate labels.

The contributions of this work are summarised as follows: **1)** To the best of our knowledge, VDDR is the first model-agnostic refiner in the field of 3D medical image segmentation which can be used for segmentation label refinement in numerous situation. **2)** We design the Dilating& Selecting framework to correct noisy labels without requiring accurate labels in practical scenarios, resulting in a **90.2%** reduction in time costs compare to manual refinement. **3)** Extensive experiments show that VDDR can effectively enhance label quality. **4)** For the first time, two new staple segmentation datasets, including noisy labels and GT are released, and code is available at https://github.com/Flq2002/3dOssSeg.

2 Method

2.1 Problem Set-up and Method Overview

In this work, we consider the problem of designing a segmentation framework to correct the noisy labels acquired from human annotators. Specifically, we consider a scenario where we have a set of 3D medical images which are divided into two groups $\{x_n^i \in \mathbb{R}^{H\times W\times D}\}_{n=1}^{N_I}$ and $\{x_n^c \in \mathbb{R}^{H\times W\times D}\}_{n=1}^{N_C}$ (with H, W, D denoting the width, height, and depth of the 3D images, and N_I, N_C representing the sample number of each group) based on different annotation types.

Table 1. Descriptions about key signs

Signs	Descriptions	Signs	Descriptions
$\tilde{y}^i$	Incomplete-structure label	$\tilde{y}^c$	Complete-structure but inaccurate label
x^i	Input image with $\tilde{y}^i$	x^c	Input image with $\tilde{y}^c$
y^i	Accurate label with x^i	y^c	Accurate label with x^c
$\tilde{x}^i$	Augmented image on x^i	$\tilde{e}^c$	Erosion result on $\tilde{y}^c$

One group consists of *incomplete-structure* labels $\{\tilde{y}_n^i \in \{0,1\}^{H\times W\times D}\}_{n=1}^{N_I}$, while the other group contains *complete-structure* but *inaccurate* labels $\{\tilde{y}_n^c \in \{0,1\}^{H\times W\times D}\}_{n=1}^{N_C}$.

To refine the *incomplete-structure* label $\tilde{y}^i$, we design a **Dilating&Selecting (D&S)** framework which mainly contains a dilating Volumetric Discrete Diffusion Refiner (VDDR) and completeness selection strategy as depicted in Figure 2. An intuitive approach is to train dilating VDDR on *complete-structure* label $\tilde{y}^c$, enabling the model to have completion capabilities. However, due to the absence of GT, overfitting the *complete-structure* label $\tilde{y}^c$ may lead to segmenting excess parts as $\tilde{y}^c$ is *inaccurate*. At the same time, insufficient training may result in an inability to complete the segmentation. To address this, we have designed a completeness selection strategy to select the most appropriate model parameters to complete the *incomplete-structure* label $\tilde{y}^i$. After post-processing, the accurate label y^i is obtained. After that, we refine inaccurate labels $\tilde{y}^c$ through training VDDR. A noise-based augmentation is designed to reduce the CT image quality disparity between x^i and x^c, yielding augmented images $\tilde{x}^i$. After training VDDR using $\tilde{x}^i$ or x^i and y^i, the model is applied on $\tilde{y}^c$ to infer the final label y^c.

2.2 Conditional Volumetric Discrete Diffusion Refiner

We proposed our Conditional Volumetric Discrete Diffusion Refiner (VDDR) based on Segrefiner [22]. In the forward process, we permute the target mask m_0, transforming it into a coarse mask m_T. The intermediate mask m_t ($t \in \{1, 2, ..., T-1\}$) is a transitional phase between m_0 and m_T which is obtained by the transition sample module proposed in Segrefiner [22]. In the reverse process, we train a conditional UNet to predict the fine mask $\tilde{m}_{0|t}$ at each timestep t. We define $s_0^{i,j,k} = [1, 0]$ and $s_T^{i,j,k} = [0, 1]$ as one-hot vectors to represent the fine and coarse state of pixel (i, j, k) in m_t, respectively. The forward process and reverse process are described as:

$$q(s_t^{i,j,k} \mid s_{t-1}^{i,j,k}) = s_{t-1}^{i,j,k} Q_t \tag{1}$$

$$p_\theta(s_{t-1}^{i,j,k} \mid s_t^{i,j,k}) = s_t^{i,j,k} P_{\theta,t}^{i,j,k} \tag{2}$$

where Q_t is a states-transition matrix and $P_{\theta,t}^{i,j,k}$ is a reversed states-transition matrix (see [22] for detailed).

Since CT images are single-channel, they lack some information compared to natural images, which are typically three-channel. To enable the model to better focus on and extract image features, instead of directly concatenating the image and mask along the channel dimension as done in [22,9,4], we designed an image encoder as illustrated in the figure 2, the number and size of the image encoders are identical to those of the UNet encoder. After obtaining multi-scale image features, we add them pixel-wise to the corresponding mask features of the same resolution to generate the final feature for UNet decoder.

2.3 Dilating&Selecting Framework

Dilating VDDR Based on VDDR, in the forward process, we gradually erode the *complete-structure* but *inaccurate* noisy label $\tilde{y}^c$ (which is more appropriately referred to as the mask in this situation), transiting it into a coarse mask $\tilde{e}^c$. In other words, we have $m_0 = \tilde{y}^c$ and $m_T = \tilde{e}^c$. In addition to the commonly used binary cross-entropy loss and dice loss, the gradient loss [5] is also introduced, which is characterized as an L1 loss between the segmentation gradient magnitudes of the predicted mask $\hat{m}_{0|t}$ and the original mask m_0. To maximize the completion capability of the model, we design a novel cover loss to constrain further $\hat{m}_{0|t}$ to be a superset of the intermediate mask m_t. The cover loss is defined as follows:

$$\mathcal{L}_{cover} = \sum(\text{ReLU}(m_t - \hat{m}_{0|t})) \tag{3}$$

where $\text{ReLu}(x) = \max(x, 0)$ and the operator $\sum$ represents the summation over all pixel values. Therefore, the objective function for training this model is:

$$\mathcal{L} = \mathcal{L}_{bce} + \mathcal{L}_{dice} + \alpha\mathcal{L}_{grad} + \beta\mathcal{L}_{cover} \tag{4}$$

where α and β are hyperparameters to balance the magnitude of all losses.

Completeness Selection Strategy To select the most appropriate model parameters, we first need to evenly sample n_p model parameters $P_{m_1}, P_{m_2}, ...P_{m_{n_p}}$ from the training process of the dilating VDDR, and apply them to the noisy labels $\tilde{y}_n^i, n = 1, 2, ..., N_I$ to obtain candidates $y_n^{m_1}, y_n^{m_2}, ..., y_n^{m_{n_p}}$, where $m_k = \lfloor \frac{N_{tot}}{n_p} \rfloor \times k$ $(k = 1, 2, ..., n_p)$ represents the iteration step and N_{tot} means the total number of training steps. Then, we train a classification model $\mathcal{C}$ to distinguish between $\tilde{y}^i$ and $\tilde{y}^c$. The model $\mathcal{C}$ is capable of discerning the completeness of the mask, outputting the probability of label completeness for each input mask (where a probability closer to 0 indicates greater incompleteness and a higher probability indicates greater completeness). By calculating the probability for each $y_n^{m_k}$ and averaging them, we obtain $Prob_{m_k} = \frac{1}{N_I}\sum_{n=1}^{N_I}\mathcal{C}(y_n^{m_k})$, $k = 1, 2, ..., n_p$. Since the dilating VDDR increasingly fits the data, the probabilities $Prob_{m_k}$ should gradually increase with the increase of m_k. We hypothesize that a sudden increase in probability, corresponding to a model where the labels are complete, indicates the desired parameter model.

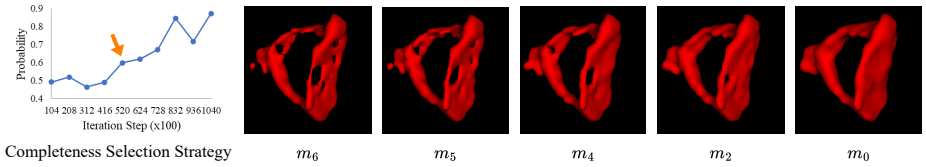

Fig. 3. The probability changes as the iteration progresses, with the orange arrows indicating the initial significant increases in value and an example of the dilating VDDR inference process based on the selected parameter. m_6 represents the coarse mask, which is progressively refined by the dilating VDDR until we obtain the fine mask m_0.

2.4 Noise-based Augmentation

We apply the NLMeans denoising [2] on the CT images x^i and x^c to obtain the estimated denoised images x^i_d, x^c_d, and the estimated noise n^i, n^c. This results in an augmented image $\tilde{x}^i = x^i_d + n^c$, which has a noise pattern similar to x^c.

3 Experiments and Results

3.1 Datasets

The temporal bone CT datasets, **OSS-I** and **OSS-C**, used in this study were collected through unilateral scanning of the temporal bone with a U-HRCT [25] scanner. The scanning parameters were set to a voltage range of 100-110 kVp and a current range of 120-180 mAs, with a reconstruction field of view of 65 mm $\times$ 65 mm. The slice thickness and interval were set to 0.1 mm, and a total of 370 slices were acquired. The images were obtained after a 20-second scan, resulting in a final image size of 370 $\times$ 650 $\times$ 650. All data were anonymized. Labeling was performed using Mimics 19.0[†], with different annotation training (as described in the Introduction), leading to label counts of 51 and 52, respectively. As this study focuses on the stapes, the region surrounding the stapes was cropped to create a 64 $\times$ 64 $\times$ 64 volume for input into the model.

3.2 Implementation Details

VDDR is trained with AdamW (lr=5×10^{-5}, batch size=1), timestep T=6, loss weight α=5, and β=0.8 for the dilating variant. Post-processing uses kernel-based sliding windows and morphological closing for output refinement. The classification model adopts Adam (lr=0.001, batch size=16, 100 epochs) with cross-entropy loss.

All experiments are implemented in PyTorch 1.12.1 and MONAI 1.3.2 on Python 3.8.19, using an Intel i9-13900K CPU and NVIDIA RTX 4090 GPUs (24 GB).

[†] https://www.materialise.com/en/healthcare/mimics/mimics-core

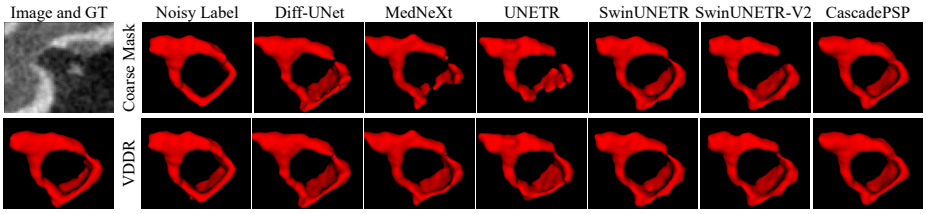

Fig. 4. Segmentation results. The first row shows the results from the baseline, while the second row presents the refined results from our VDDR.

3.3 Refinement of OSS-I

In the D&S framework used to refine OSS-I, we set $n_p = 10$ and $N_{tot} = 104,000$. The variation in probabilities, $Prob_{m_k}$ for $k = 1, 2, \ldots, 10$, is shown in Figure 3. As seen, the most significant change in probabilities occurs at the 52,000th step, which is halfway through the training process. This aligns with the idea that a small step size may lead to underfitting, while a large step size would cause overfitting. Therefore, we chose the model parameters from this step for inference. An example of the inference process is shown in Figure 3, where the model progressively refines the incomplete areas within the mask.

The inference results were then submitted to an expert for evaluation based on the following criteria: A noisy label indicating over-segmentation is scored 0. If the noisy label represents under-segmentation, Score 1 means the refined label is unreasonable; Score 2 means it is generally reasonable; and Score 3 means the refined label is completely accurate. After excluding two cases that received Score 0, the proportion of refined labels rated at least Score 2 was **100%**, with **73.5%** (36/49) rated as completely accurate. This demonstrates the practical effectiveness of the proposed framework.

Additionally, we asked the expert to refine the remaining labels using the model's outputs, significantly reducing the time compared to correcting the original noisy labels directly. For experts experienced in annotation, the former approach took an average of only **10 minutes** per revision, compared to **30 minutes** for direct corrections. This led to a **90.2%** reduction in time spent on annotating our OSS-I dataset.

3.4 Refinement of OSS-C

After obtaining the accurate labels for OSS-I, a direct approach to obtaining accurate labels for OSS-C would be to train the VDDR using OSS-I and directly infer on OSS-C to achieve expert-satisfactory labels. However, in practice, due to limited, low-quality CT image and the higher difficulty of label refinement for OSS-C compared to OSS-I (the latter only requiring dilation refinement), the expert satisfaction rate was less than 10%. Therefore, we asked experts to manually correct the labels of OSS-C, setting OSS-I as the training set and OSS-C as the test set to further investigate the performance of our VDDR.

Table 2. Evaluate other segmentation methods and apply our VDDR to their results. The results in **bold** represent the optimal outcomes of our model, while the underlined results indicate the optimal outcomes of the baseline.

Methods	Coarse Mask				VDDR (ours)			
	Dice↑	Jaccard↑	95HD↓	ASD↓	Dice↑	Jaccard↑	95HD↓	ASD↓
MedNeXt [17]	69.92	54.09	2.62	0.55	74.07	59.04	2.36	0.55
Diff-UNet [23]	73.31	58.10	3.86	0.90	73.94	58.91	4.02	0.92
UNETR [11]	69.21	53.31	2.66	0.67	73.07	57.83	2.53	0.60
SwinUNETR [10]	75.73	61.31	2.49	0.56	76.70	62.50	2.37	0.54
SwinUNETR-V2 [12]	76.39	62.10	2.30	0.53	77.09	62.99	2.20	0.51
CascadePSP [5]	75.85	61.32	2.09	0.47	76.24	61.80	**2.07**	**0.46**
Noisy Label (manually)					**77.14**	**63.18**	2.40	0.59
w/o noise-based aug	70.23	54.79	3.43	0.66	75.31	61.05	2.95	0.57
w/o image encoder					73.05	58.09	3.14	0.59

Five state-of-the-art (SOTA) 3D medical image segmentation methods are chosen for comparison, including one CNN-based method (MedNeXt [17]), one diffusion-based method (Diff-UNet [23]) and three transformer-based methods (UNETR[11], SwinUNETR [10], and SwinUNETR-V2 [12]). We also modified the classic 2D refiner network CascadePSP [5] into a format capable of accepting 3D image inputs for further comparative analysis. We choose four evaluation metrics: Dice Score (%), Jaccard Score (%), 95% Hausdorff Distance (95HD) in voxel and Average Surface Distance (ASD) in voxel.

Comparison with SOTA methods The quantitative results are reported in Table 2. The coarse mask on the left side represents the predictions of different methods. Specifically, the prediction results of CascadePSP are based on noisy label, which refers to inaccurate labels from manual annotation. The right side shows the results after refinement by our model. It can be observed that, except for Diff-UNet, the performance metrics of the other methods have improved to varying degrees by our VDDR. When the coarse mask is the noisy label, our model achieves optimal results in terms of Dice and Jaccard scores. The visualization results are shown in the Figure 4, where the corrected outcomes show improvements while retaining the structural features of the coarse masks.

Ablation Studies The results of our ablation experiments using noisy label as the coarse mask are presented in the last two rows of the Table 2 and "w/o image encoder" indicates that the image and mask are concatenated along the channel dimension and input into a UNet with 2 input-channel. It is evident that both noisy augmentation and the image encoder significantly enhance the model's performance.

4 Conclusions

This paper addresses noisy labels in ossicular chain segmentation and proposes a label refinement method. The Conditional Volumetric Discrete Diffusion Refiner (VDDR), which combines an image encoder with a discrete diffusion process, effectively refines segmentation labels. The Dilating&Selection framework, using a dilating VDDR with cover loss and completeness selection, corrects incomplete labels without accurate annotations. This framework could provide valuable insights for other applications. VDDR consistently refines coarse labels under various conditions, demonstrating robustness. Future work will explore large-scale dataset applications.

Acknowledgments. This work was supported by the National Key R&D Program of China (2020YFA0712203), National Natural Science Foundation of China (62371316, 82302282, 62276012), Beijing Science and Technology Plan Project (Z241100009024020), Beijing Scholar 2015 ([2015]160), and Capital's Funds for Health Improvement and Research (2022-1-1111).

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

References

1. Austin, J., Johnson, D.D., Ho, J., Tarlow, D., Van Den Berg, R.: Structured denoising diffusion models in discrete state-spaces. Advances in neural information processing systems **34**, 17981–17993 (2021)
2. Buades, A., Coll, B., Morel, J.M.: A non-local algorithm for image denoising. In: 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05). vol. 2, pp. 60–65. Ieee (2005)
3. Chen, T., Wang, C., Chen, Z., Lei, Y., Shan, H.: Hidiff: hybrid diffusion framework for medical image segmentation. IEEE Transactions on Medical Imaging (2024)
4. Chen, T., Wang, C., Shan, H.: Berdiff: Conditional bernoulli diffusion model for medical image segmentation. In: International conference on medical image computing and computer-assisted intervention. pp. 491–501. Springer (2023)
5. Cheng, H.K., Chung, J., Tai, Y.W., Tang, C.K.: CascadePSP: Toward class-agnostic and very high-resolution segmentation via global and local refinement. In: CVPR (2020)
6. Dahmani-Causse, M., Marx, M., Deguine, O., Fraysse, B., Lepage, B., Escudé, B.: Morphologic examination of the temporal bone by cone beam computed tomography: comparison with multislice helical computed tomography. European annals of otorhinolaryngology, head and neck diseases **128**(5), 230–235 (2011)
7. Fang, C., Wang, Q., Cheng, L., Gao, Z., Pan, C., Cao, Z., Zheng, Z., Zhang, D.: Reliable mutual distillation for medical image segmentation under imperfect annotations. IEEE Transactions on Medical Imaging **42**(6), 1720–1734 (2023)
8. Fauser, J., Stenin, I., Bauer, M., Hsu, W.H., Kristin, J., Klenzner, T., Schipper, J., Mukhopadhyay, A.: Toward an automatic preoperative pipeline for image-guided temporal bone surgery. International journal of computer assisted radiology and surgery **14**, 967–976 (2019)

9. Guo, X., Yang, Y., Ye, C., Lu, S., Peng, B., Huang, H., Xiang, Y., Ma, T.: Accelerating diffusion models via pre-segmentation diffusion sampling for medical image segmentation. In: 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI). pp. 1–5. IEEE (2023)
10. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In: International MICCAI brainlesion workshop. pp. 272–284. Springer (2021)
11. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D.: Unetr: Transformers for 3d medical image segmentation. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 574–584 (2022)
12. He, Y., Nath, V., Yang, D., Tang, Y., Myronenko, A., Xu, D.: Swinunetr-v2: Stronger swin transformers with stagewise convolutions for 3d medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 416–426. Springer (2023)
13. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
14. Ivanovic, A., Schalbetter, F., Schmeltz, M., Wimmer, W., Caversaccio, M., Stampanoni, M., Bonnin, A., Anschuetz, L.: Characterizing bone density pattern and porosity in the human ossicular chain using synchrotron microtomography. Scientific reports **14**(1), 18498 (2024)
15. Neves, C., Tran, E., Kessler, I., Blevins, N.: Fully automated preoperative segmentation of temporal bone structures from clinical ct scans. Scientific reports **11**(1), 116 (2021)
16. Nikan, S., Van Osch, K., Bartling, M., Allen, D.G., Rohani, S.A., Connors, B., Agrawal, S.K., Ladak, H.M.: Pwd-3dnet: a deep learning-based fully-automated segmentation of multiple structures on temporal bone ct scans. IEEE Transactions on Image Processing **30**, 739–753 (2020)
17. Roy, S., Koehler, G., Ulrich, C., Baumgartner, M., Petersen, J., Isensee, F., Jaeger, P.F., Maier-Hein, K.H.: Mednext: transformer-driven scaling of convnets for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 405–415. Springer (2023)
18. Shi, J., Zhang, K., Guo, C., Yang, Y., Xu, Y., Wu, J.: A survey of label-noise deep learning for medical image analysis. Medical Image Analysis **95**, 103166 (2024)
19. Shin, H.Y., Hwang, H.J.: Mental health of the people with hearing impairment in korea: A population-based cross-sectional study. Korean journal of family medicine **38**(2), 57 (2017)
20. Song, H., Kim, M., Park, D., Shin, Y., Lee, J.G.: Learning from noisy labels with deep neural networks: A survey. IEEE Transactions on Neural Networks and Learning Systems **34**(11), 8135–8153 (2023)
21. Tang, R., Zhao, P., Li, J., Wang, Z., Xu, N., Wang, Z.: Artificial intelligence in ct diagnosis: Current status and future prospects for ear diseases. Meta-Radiology **2**(4), 100112 (2024)
22. Wang, M., Ding, H., Liew, J.H., Liu, J., Zhao, Y., Wei, Y.: SegRefiner: Towards model-agnostic segmentation refinement with discrete diffusion process. In: NeurIPS (2023)
23. Xing, Z., Wan, L., Fu, H., Yang, G., Zhu, L.: Diff-unet: A diffusion embedded network for volumetric segmentation (2023)
24. Xu, Z., Lu, D., Luo, J., Wang, Y., Yan, J., Ma, K., Zheng, Y., Tong, R.K.Y.: Anti-interference from noisy labels: Mean-teacher-assisted confident learning for medical

image segmentation. IEEE Transactions on Medical Imaging **41**(11), 3062–3073 (2022)
25. Yin, H., Zhao, P., Zhang, L., Lv, H., Wang, Z., Zhang, P., et al.: An experimental study on the ability of newly developed ct equipment for temporal bone to display fine bony anatomy. Chin J Radiol **54**, 763–768 (2020)