From Variability To Accuracy: Conditional Bernoulli Diffusion Models with Consensus-Driven Correction for Thin Structure Segmentation

Jinseo An[1][0000−0003−0919−8458], Min Jin Lee[1][0000−0002−6773−1364], Kyu Won Shim[2][0000−0002−9441−7354], and Helen Hong[1⋆][0000−0001−5044−7909]

[1] Department of Software Convergence, Seoul Women's University, Seoul, Republic of Korea
{jsan,minjin,hlhong}@swu.ac.kr

[2] Department of Pediatric Neurosurgery, Craniofacial Reforming and Reconstruction Clinic, Yonsei University College of Medicine, Seoul, Republic of Korea
shimkyuwon@yuhs.ac

Abstract. Accurate segmentation of orbital bones in facial computed tomography (CT) images is essential for the creation of customized implants for reconstruction of defected orbital bones, particularly challenging due to the ambiguous boundaries and thin structures such as the orbital medial wall and orbital floor. In these ambiguous regions, existing segmentation approaches often output disconnected or under-segmented results. We propose a novel framework that corrects segmentation results by leveraging consensus from multiple diffusion model outputs. Our approach employs a conditional Bernoulli diffusion model trained on diverse annotation patterns per image to generate multiple plausible segmentations, followed by a consensus-driven correction that incorporates position proximity, consensus level similarity, and gradient direction similarity to correct challenging regions. Experimental results demonstrate that our method outperforms existing methods, significantly improving recall in ambiguous regions while preserving the continuity of thin structures. Furthermore, our method automates the manual process of segmentation result correction and can be applied to image-guided surgical planning and surgery.

Keywords: Segmentation · Diffusion model · Consensus · Correction · Inter-observer variability.

1 Introduction

Orbital bone fractures commonly occur in thin regions, such as orbital medial wall and orbital floor [6]. Accurate orbital bone segmentation in computed tomography (CT) images is crucial in craniomaxillofacial surgery, particularly for

⋆ Corresponding author

designing patient-specific implants and establishing image-guided surgical plans. However, segmenting these thin bone structures presents significant challenges due to their low contrast with surrounding tissues and ambiguous boundaries caused by partial volume effects in thin structures [10], leading to inter-observer variability in manual annotations. Previous study has quantified this variability using intra-class correlation coefficient (ICC) in manual orbital bone annotation, reporting lower consensus in thin bone regions (ICC=0.715 for orbital medial wall, ICC=0.824 for orbital floor) compared to whole orbital bone (ICC=0.931) [9]. Despite recent attempts such as MSDA-Net [3], which applied multi-scale and dual attention modules to improve segmentation accuracy for orbital bones of varying thickness, evaluation results showed variation depending on which annotation was used as reference standard [2].

Recently, diffusion models have shown remarkable advances in medical segmentation tasks [16,7]. Unlike traditional CNN-based methods that produce deterministic results, diffusion models leverage the stochastic nature of noise sampling to generate diverse plausible segmentations. We argue that this inherent capability can provide insight when dealing with particularly thin structures and ambiguous boundaries that even expert annotators exhibit significant variability. Although this stochastic nature enables the generation of diverse segmentations, producing meaningful segmentation masks requires conditioning on the corresponding input image to guide the generation process [8]. Conditional diffusion models improve segmentation performance by incorporating anatomical structure information from medical images as conditioning input [1,11]. While most existing studies rely on Gaussian noise, several studies have proposed using Bernoulli noise instead, which is more appropriate for binary mask segmentation tasks due to the discrete nature of the masks [5,13].

In this paper, we present a novel framework that combines diffusion model-based segmentation with consensus-driven correction to improve accuracy in challenging regions. Our main contributions are: (1) We employ a conditional Bernoulli diffusion model for segmentation, providing three different annotation masks per input image to learn inter-observer variability. (2) We propose consensus-driven correction to address the inherent variation in ambiguous boundaries, considering position proximity, consensus level similarity, and gradient direction similarity across multiple segmentations from the diffusion model. (3) Experimental results on thin bones of orbital medial wall and orbital floor demonstrate that our method corrects challenging regions and segmentation performance outperforms other comparison methods.

2 Method

2.1 Conditional Bernoulli Diffusion Model for Image Segmentation

Fig. 1 represents the overview of our conditional Bernoulli diffusion model for learning diverse annotation patterns and generating multiple plausible segmentations. Our implementation is based on BerDiff [4], a diffusion model based on

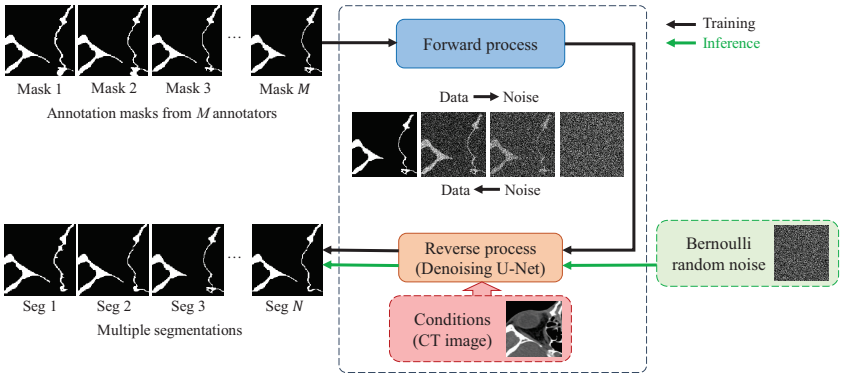

Fig. 1. Illustration of our conditional Bernoulli diffusion model

Bernoulli noise, that is more appropriate for binary medical image segmentation tasks, where masks consist exclusively of discrete values (0 or 1).

The model takes M annotations from multiple annotators per input image to capture inter-observer variability. During the forward process, binary masks adds noise through Bernoulli noise sampling. In the reverse process, the model progressively removes noise to generate segmentations, using CT images as a condition to guide the generation process. Conditioning is implemented by concatenating CT images with the noisy masks as input to the denoising network, providing intensity and anatomical context information that enables the model to learn structural patterns during the denoising.

An advantage of our approach is the ability to learn from various annotation patterns, reducing dependency on subjective annotation by any single annotator. By training on M different annotations per input image, the model captures the inherent inter-observer variability in challenging regions. This is particularly valuable for ambiguous boundaries where even expert annotators disagree. In these regions, our diffusion model generates varying plausible segmentations based on the learned annotation distribution, which we identify as challenging areas requiring consensus-driven correction.

To capture the level of consensus among plausible segmentations, we generate a consensus-driven uncertainty map by aggregating multiple segmentations and normalizing them to a [0, 255] range, as shown in Fig. 2. This map provides a spatial representation of uncertainty, with higher values indicating stronger consensus level across segmentations and lower values highlighting regions of uncertainty. This uncertainty information serves as a foundation for our subsequent consensus-driven correction method.

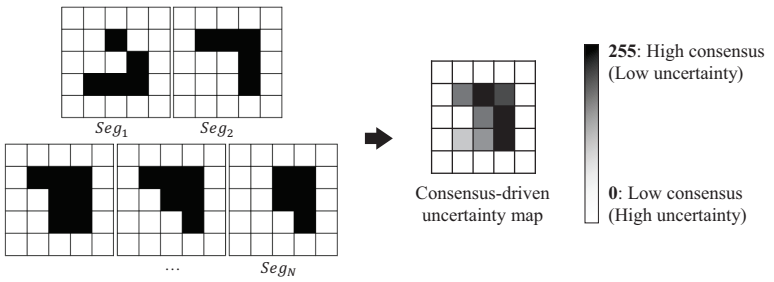

Fig. 2. Process of generating a consensus-driven uncertainty map

2.2 Consensus-driven Correction

We are dealing with ambiguous regions that are inherently unclear in CT images, so low consensus (high uncertainty) exhibits in the consensus-driven uncertainty map. Even with low consensus, nearby pixels with similar spatial and consensus characteristics are identified as potential pixels for correction. We leverage the consensus-driven uncertainty map, which provides valuable insight into potentially correct segmentation areas.

Consensus-driven correction minimizes the energy function in Eq. 1, which combines the unary potential (Eq. 2) and pairwise potential (Eq. 3). The unary potential is derived from the consensus level from the map, providing a pixel-wise measure of segmentation confidence. The pairwise potential incorporates spatial and consensus information from the map, including position proximity, consensus level similarity, and gradient direction similarity. Position proximity helps maintain consistent segmentation by considering close regions, which is important for preserving anatomical continuity in thin structures. Consensus level similarity ensures that similar uncertain regions are corrected consistently. Gradient direction similarity helps maintain structural directional patterns, which is critical for thin structure to preserve shape. For example, even when consensus is low in a specific region, our method may correct a pixel classification based on its proximity to a neighboring pixel that is segmented as bone and has similar consensus level and gradient direction. The energy function is formulated as follows:

$$E(x) = \sum_i \psi_i(x_i) + \sum_{ij} \psi_{ij}(x_i, x_j) \tag{1}$$

$$\psi_i(x_i) = -logP(x_i) \tag{2}$$

$$\psi_{ij}(x_i, x_j) = \mu(x_i, x_j)[w_1 \exp(-\frac{|p_i - p_j|^2}{2\theta_\alpha^2} - \frac{|c_i - c_j|^2}{2\theta_\beta^2} - \frac{|d_i - d_j|^2}{2\theta_\gamma^2}) + w_2 \exp(-\frac{|p_i - p_j|^2}{2\theta_\delta^2})] \tag{3}$$

where x_i, x_j represent the class of pixel i,j. $P(x_i)$ in the unary potential denotes the probability of pixel i belongs to a specific class, as derived from the consensus-driven uncertainty map. The pairwise potential contains three Gaussian kernels accounting for positional proximity (p), consensus level similarity (c), and gradient direction similarity (d). The class compatibility function $\mu(x_i, x_j)$ is defined based on the Potts model: $\mu = 1$ if $x_i \neq x_j$, $\mu = 0$ otherwise. This penalizes assigning different classes to nearby, similar pixels. The parameters $\theta_\alpha, \theta_\beta, \theta_\gamma$, and θ_δ are scaling factors for the Gaussian kernels, controlling the influence range of each characteristic. These scaling factors were determined through multiple experiments considering the image size, consensus level range, and gradient direction range. We set $\theta_\alpha = 80, \theta_\beta = 60, \theta_\gamma = 2, \theta_\delta = 3, w_1 = 15, w_2 = 1$.

3 Experiment

3.1 Dataset and Preprocessing

This study was approved by the Institutional Review Board of Severance Hospital, Yonsei University College of Medicine, Seoul, Republic of Korea (IRB No. 4-2016-0603). The dataset comprises facial CT images from 71 patients, divided into a training set of 57 cases and a test set of 14 cases. All images have a matrix size of 512×512 pixels, with in-plane resolution ranging from 0.4 to 0.619mm and a slice thickness of 1mm. Each image was manually annotated by three annotators—a neurosurgeon with over 15 years of experience and two senior medical students—following the same annotation protocols.

We performed preprocessing to ensure consistency across different CT scanners and acquisition protocols. This included intensity normalization using a window width of 600 HU and window level of 100 HU, followed by conversion from 12-bit to 8-bit representation (0-255 intensity range). Additionally, all images were resampled to a uniform pixel spacing of 0.4 x 0.4 mm^2, corresponding to the highest resolution present in the dataset.

3.2 Experimental Setup and Implementation Details

Comparison Methods. We compare our method with both CNN-based and diffusion-based segmentation approaches. The CNN-based methods include U-Net [12] and MSDA-Net [3], while diffusion-based methods include MedSegDiff-v2 (Gaussian noise) [15] and BerDiff (Bernoulli noise) [4]. CNN-based methods produce deterministic results for a given trained model, whereas diffusion-based methods generate multiple segmentations from a single model due to their stochastic nature.

Evaluation Metrics. Three metrics are used for performance evaluation, including Dice Similarity Coefficient (DSC), recall, and precision. As reference standard, we use masks generated by the STAPLE algorithm [14] from three annotations, which is widely used approach to combine multiple expert annotations

Table 1. Segmentation performance of our proposed method and comparison methods. The best results are highlighted in bold. '*': p<0.001 compared to our method based on a t-test.

(%)

Methods	Orbital medial wall			Orbital floor		
	DSC	Recall	Precision	DSC	Recall	Precision
U-Net [12]	82.28*	79.70*	85.92	89.86	90.24*	89.85
MSDA-Net [3]	83.39*	84.18*	83.26*	89.99	92.72	87.79*
MedSegDiff-v2 [15]	84.07*	81.58*	87.26*	90.11*	88.88*	**92.02***
BerDiff [4]	86.14	84.60*	**88.16***	**91.53**	92.15	91.37
BerDiff + Correction (Ours)	**86.31**	**87.83**	85.38	91.36	**93.24**	90.08

to generate a reference standard. To focus on thin bone regions, we manually define two evaluation ROIs for orbital medial wall and orbital floor.

Implementation Details. The diffusion model is trained with a batch size of 2, using the AdamW optimizer with a learning rate of 5e-5 and a linear noise schedule over 1000 timesteps. The diffusion model is trained for 100,000 iterations on a NVIDIA GeForce RTX 3090 GPU using Python 3.7 and PyTorch 1.11. During training, one of multiple annotations was randomly selected for each iteration, allowing the model to learn various annotation patterns. We adopted the DDIM sampling strategy for BerDiff and the DDPM sampling strategy for MedSegDiff-v2, following their original implementations, and generated 200 segmentations for each diffusion-based method. For evaluation, CNN-based methods produced a single segmentation result, whereas diffusion-based methods generated 200 segmentations, which were averaged to obtain the final output. Our method also utilized 200 segmentations from BerDiff for correction.

3.3 Results

We present the quantitative and qualitative results in Table 1 and Fig. 3, respectively. As can be seen in Table 1, our method outperforms all CNN-based methods in all metrics, including DSC, recall, and precision. In particular, the recall of the orbital medial wall showed statistically significant improvement with $p < 0.001$. This improvement is particularly important for addressing under-segmentation at ambiguous boundaries while preserving continuity in thin structures.

Fig. 3 demonstrates our correction capabilities in the challenging regions. The red boxes highlight ambiguous regions, which show high uncertainty in the consensus-driven uncertainty map. CNN-based methods show disconnected segmentation in those regions. BerDiff shows improved results compared to CNN-based methods, but still fails, leaving disconnected regions. However, our proposed method successfully corrects thin structures in these challenging areas by leveraging the clear vertical directionality of the orbital medial wall through a consensus-driven correction approach.

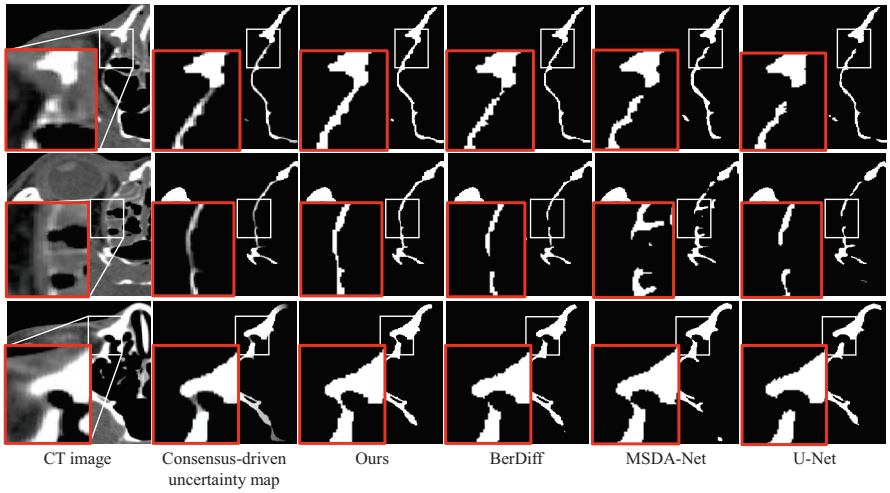

Fig. 3. Representative orbital bone segmentation results. Consensus-driven uncertainty map is generated by accumulating multiple segmentations using a conditional Bernoulli diffusion model (BerDiff). The red box is zoomed in to emphasize the thin region within the white box.

4 Conclusion

In this paper, we proposed a framework combining conditional Bernoulli diffusion model-based segmentation with consensus-driven correction. This approach aims to improve the segmentation performance of thin structures in orbital bones, with a particular focus on the orbital medial wall and orbital floor, which possess ambiguous boundaries. In our experiments, we addressed under-segmentation issues in ambiguous regions by leveraging the ability of our proposed method to learn from various annotation patterns. This approach reduces dependency on subjective annotation by any single annotator and enables us to use the consensus among multiple segmentations from the diffusion model as valuable information. By considering directionality in our consensus-driven correction method, we achieved improvements of up to 4.03% in DSC and 8.13% in recall for orbital medial wall compared to CNN-based methods. Our novel framework learns data distribution from multiple annotations in ambiguous regions and ensures consistent segmentation results while reducing manual correction process, making it applicable to surgical planning and customized implant creation.

Acknowledgments. This research was partially supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HI22C1496) and the National Research Foundation of Korea (NRF) grants funded by the Korea government (MSIT) (RS-2024-00340610 and RS-2023-00207947).

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

References

1. Amit, T., Shaharbany, T., Nachmani, E., Wolf, L.: Segdiff: Image segmentation with diffusion probabilistic models. arXiv preprint (2021), https://arxiv.org/abs/2112.00390
2. An, J., Lee, M.J., Hong, H., Shim, K.W.: Evaluation of segmentation performance using multiple reference standards for accurate orbital bone modeling in craniomaxillofacial surgery: based on msda-net deep learning segmentation. Presented at Radiological Society of North America (RSNA) (2023)
3. An, J., Lee, M.J., Shim, K.W., Hong, H.: Orbital bone segmentation using improved skip connection of u-net structure in facial ct images. Journal of the Korea Computer Graphics Society **29**(2), 13–20 (2023). https://doi.org/10.15701/kcgs.2023.29.2.13
4. Chen, T., Wang, C., Shan, H.: Berdiff: Conditional bernoulli diffusion model for medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2023. Lecture Notes in Computer Science, vol. 14223, pp. 491–501. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-43901-8_47
5. Chen, T., Wang, C., Chen, Z., Lei, Y., Shan, H.: Hidiff: Hybrid diffusion framework for medical image segmentation. IEEE Transactions on Medical Imaging **43**(10), 3570–3583 (2024). https://doi.org/10.1109/TMI.2024.3424471
6. Cho, R.I., Davies, B.W.: Combined orbital floor and medial wall fractures involving the inferomedial strut: repair technique and case series using preshaped porous polyethylene/titanium implants. Craniomaxillofacial trauma & reconstruction **6**(3), 161–169 (2013). https://doi.org/10.1055/s-0033-1343785
7. Chowdary, G.J., Yin, Z.: Diffusion transformer u-net for medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2023. Lecture Notes in Computer Science, vol. 14223. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-43901-8_59
8. Hejrati, B., Banerjee, S., Glide-Hurst, C., Dong, M.: Conditional diffusion model with spatial attention and latent embedding for medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. Lecture Notes in Computer Science, vol. 15009, pp. 202–212. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-72114-4_20
9. Kim, H., et al.: Comparative evaluation of inter-observer variability in manual segmentation of orbital bone from facial ct images. Presented at The Korean Society of 3D Printing in Medicine (2022)
10. Kim, H., et al.: Three-dimensional orbital wall modeling using paranasal sinus segmentation. Journal of Cranio-Maxillofacial Surgery **47**(6), 959–967 (2019). https://doi.org/10.1016/j.jcms.2019.03.028
11. Rahman, A., Valanarasu, J.M.J., Hacihaliloglu, I., Patel, V.M.: Ambiguous medical image segmentation using diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11536–11546 (2023)

12. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Lecture Notes in Computer Science, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
13. Wang, Z., Wang, J., Liu, Z., Qiu, Q.: Binary latent diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 22576–22585 (2023)
14. Warfield, S.K., Zou, K.H., Wells, W.M.: Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation. IEEE Transactions on Medical Imaging **23**(7), 903–921 (2004). https://doi.org/10.1109/TMI.2004.828354
15. Wu, J., Ji, W., Fu, H., Xu, M., Jin, Y., Xu, Y.: Medsegdiff-v2: Diffusion-based medical image segmentation with transformer. Proceedings of the AAAI Conference on Artificial Intelligence **38**(6), 6030–6038 (2024). https://doi.org/10.1609/aaai.v38i6.28418
16. Yan, P., et al.: Cold segdiffusion: A novel diffusion model for medical image segmentation. Knowledge-Based Systems **301**, 112350 (2024). https://doi.org/10.1016/j.knosys.2024.112350