Entropy-weighted Voting Method for Diffusion-based Semantic Segmentation

Seonggyun Jeong
Dept. of Electrical and Computer Engineering
Ajou University
Suwon, Korea
zkdlghzl5@ajou.ac.kr

Yong Seok Heo
Dept. of Electrical and Computer Engineering
Dept. of Artificial Intelligence
Ajou University
Suwon, Korea
ysheo@ajou.ac.kr

***Abstract*—Recently, semantic segmentation methods leveraging image generation models have garnered significant attention. In particular, approach based on Diffusion models (DDPM) that utilize mid-level activations from the diffusion network with a majority voting of distributions from several light multi-layer perceptron (MLP) have shown better performance compared to GAN-based approaches. However, utilizing a simple majority voting system is suboptimal. In this paper, we propose a novel voting method for DDPM-based semantic segmentation. Our method introduces a weighted sum of distributions, where the weights are determined by the entropy of the class prediction results obtained from each MLP model. We conduct experiments on various datasets, including LSUN-Bedroom, FFHQ-256, LSUN-Cat, and LSUN-Horse. The results demonstrate that our proposed method achieves better mean Intersection over Union (mIoU) scores compared to previous work.**

***Index Terms*—semantic segmentation, diffusion, voting, entropy**

I. Introduction

The goal of semantic segmentation is to perform pixel-wise classification, which is one of the computer vision tasks. Recent researches [1], [2], [3] have been actively performed to solve semantic segmentation tasks by extracting representation from the image generation models such as Generative Adversarial Networks (GANs). The GAN-based method [4] is one of approaches, which converts the image into latent code through the learned encoder and extracts representations (feature map) from the style-based generator to predict the segmentation map by learning ensemble of 10 per-pixel classifiers. However, this method can suffer from mode-collapse in terms of diversity in that it uses GANs, and also requires an additional encoder that maps images to latent code.

To overcome these limitations, Denoising Diffusion Probabilistic Models (DDPM) [5]-based method [6] has been explored. They train an ensemble of 10 per-pixel classifiers to predict the segmentation map in the same way as the GAN-based method [4]. Since it is based on DDPM, no additional encoder learning is required, and various representations can be obtained from the generator compared to the GAN-based method. Thus, it has better performance in terms of mean Intersection over union (mIoU) for segmentation tasks. After each classifiers training, the models used as an ensemble for semantic segmentation, and the most frequently prediction class is determined as a final prediction class using a hard-voting method. In this process, the uncertainty of the probability distribution of each per-pixel classifier is not considered, which is suboptimal.

In this paper, we propose a new voting method of using the weighted sum of the entropy of each model's probability distribution as a final probability distribution.

The contributions of the proposed method are as follows:

- Our method can improve prediction performance, because the entropy of the distribution is treated as weight, the proportion of models with a more certain probability distribution is increased.
- It is possible to form a smooth decision boundaries by switching from hard-voting to a soft-voting method.

II. Related Work

A. Semantic segmentation with generation models.

With the development of image generation models, studies applying them to downstream tasks such as semantic segmentation have become increasingly active. Among the prior works, Nontawat Tritrong *et al*. [1] has focused on extracting pixel-wise representations from trained GANs, which were used as input for segmentation models, demonstrating comparable results to supervised methods. LinearGAN [2] proposed a semantic segmentation method with trained GANs. It utilized the GAN generator's feature maps in the image generation process, where the segmentation is obtained through a series of steps involving up-sampling, concatenation, and linear transformations of the feature maps. Galeev *et al*. [3] showed that it is possible to generate segmentation map with lightweight Multi Layer Perceptron (MLP) from the representation of the GAN generator. Additionally, it proposed a structure for approximating representations for unsupervised domain-specific pre-training. By utilizing a GAN generator with an additional encoder that reconstructs to a latent code, DatasetGAN [4] proposed a structure in which the GAN generator's feature maps are passed through an ensemble of MLP classifiers for the semantic segmentation task. The DDPM-based segmentation method [6], while sharing a similar pipeline to GAN-based methods, stands apart by utilizing DDPM instead of GAN for image generation. Notably, it demonstrated improved

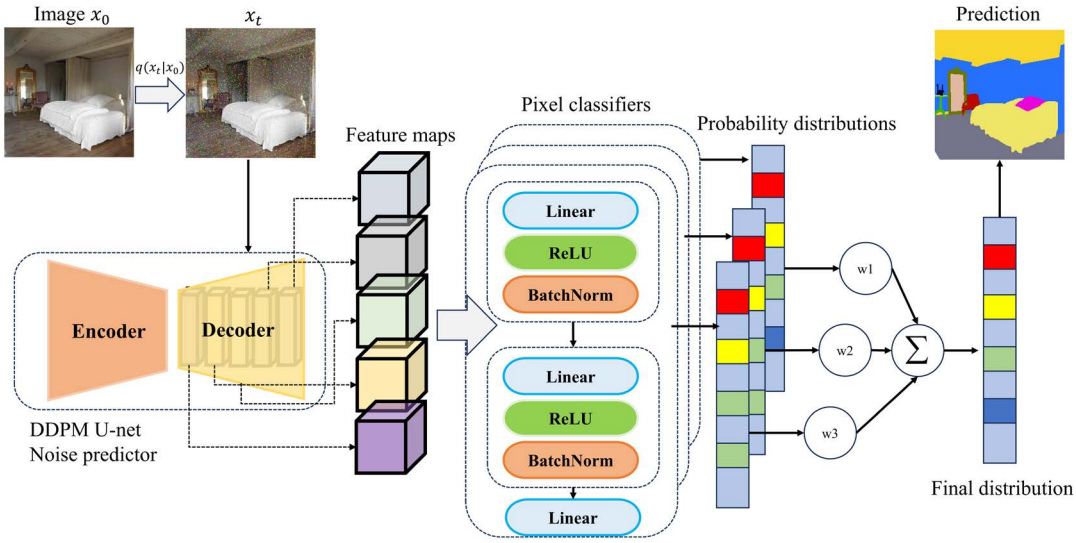

Fig. 1. **Overview of proposed method.** Utilizing feature maps of the noise predictor (DDPM U-net)'s decoder, each classifiers makes probability distribution for segmentation task. To eliminate uncertainty of distribution, we propose a weighted sum of distribution as final distribution.

segmentation performance compared to GAN-based methods without an additional encoder and yielded superior results when trained with real images. However, the final prediction is not optimized because the uncertainty of each model is not taken into consideration during the process of ensembling pixel classifiers. In this paper, we propose a way to improve the problem and compare it with previous studies.

B. *Diffusion.*

Diffusion probabilistic models [7] are image generation methods inspired by non-equilibrium statistical physics. These models predicted noise from a Gaussian distribution at a specific time-step using a Markov chain. The noise-conditioned score network [8] generated images through score-matching and learns a score network capable of estimating gradients and samples through Langevin dynamics. DDPM [5] redefined the loss function of diffusion probabilistic models to prioritize the denoising process, resulting in improved performance. Additionally, the Denoising Diffusion Implicit Models (DDIM) [9] method was modified within the DDPM approach to enable sampling through a non-Markovian process, effectively reducing the problematic sampling time that was previously encountered in DDPM. The Guided Diffusion method [10] had improved the fidelity of generated images compared to previous works and outperformed GANs by incorporating conditions into the sampling process using a classifier in the DDPM model. Ablated Diffusion Model(ADM) [10] refers to the diffusion model trained using the method mentioned earlier. For our experiments, we adopt the ADM network proposed in [10] as the generation network.

III. PROPOSED METHOD

A. *Analysis of previous works*

The train, test pipeline and network structure of DDPM-based segmentation [6] is the same as follows. After freezing the weights of the learned ADMs [10], feature maps are extracted from specific blocks of the decoder. The feature maps are bilinearly interpolated to match the resolution of the segmentation map and then concatenated. The concatenated feature map vector is input into the classifier in a supervised manner, and this process is repeated for each of the 10 pixel classifiers. In the testing process, after training the classifiers separately, predictions are generated by combining the outputs of an ensemble of pixel classifiers. There are a total of 10 classifiers, namely Model 1, Model 2, ..., Model 10. For each classifier, the feature maps of the noise predictor network are processed through the respective model. The input image set $\{X_1, X_2, \ldots, X_N\}$ is input into the classifier, producing feature maps $\{S_1, S_2, \ldots, S_N\}$ in $\mathbb{R}^{HW \times D}$. Then softmax operation is performed along dimension D on the feature maps. it means probability distribution for the class at each pixel of feature maps. Using an ensemble of classifiers, the probability distribution of each classifier assigns an index (class) to the pixel with the highest probability. From a hard voting perspective, the final prediction class is determined by selecting the most frequently chosen index among multiple classifiers. This method not only neglect the uncertainty of the distribution of each classifier, but also has non-smooth decision boundaries. To address these issues, it is necessary to transition from a hard voting method to a soft voting method in the ensemble, and also consider the uncertainty of the distribution.

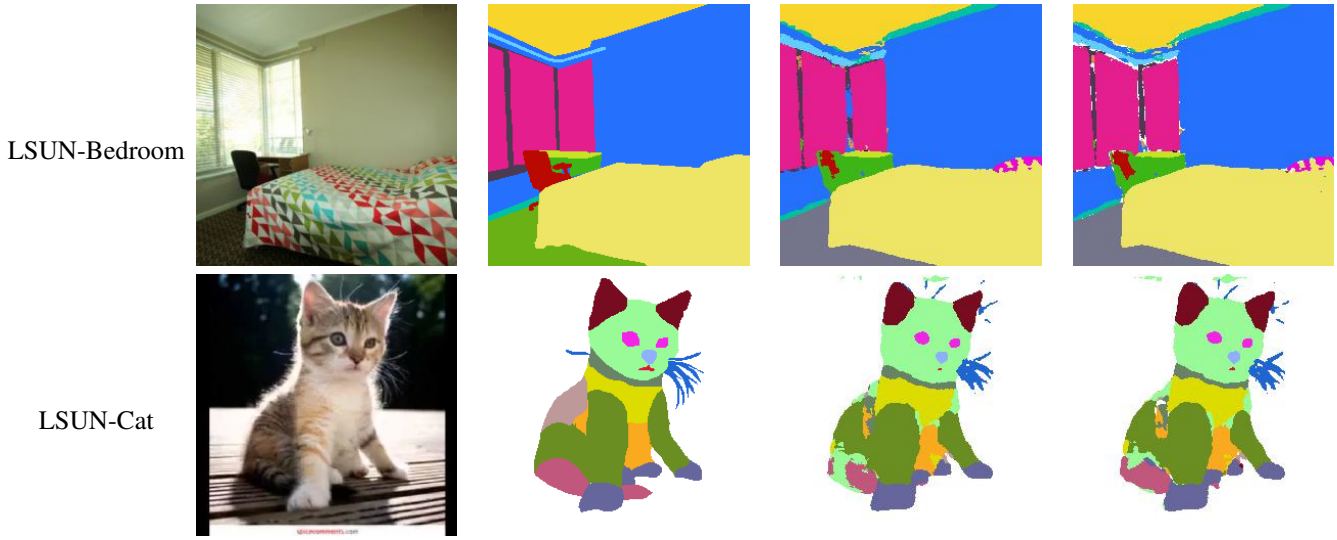

Fig. 2. **Comparisons with DDPM-based method and Ours.** Our method achieves a higher mIoU score of 0.8(%) for a specific LSUN-Bedroom image compared to the DDPM-based method, and a higher mIoU score of 2.2(%) for a specific LSUN-Cat image compared to the DDPM-based method

B. Entropy-weighted voting method

The class probability distribution of i^{th} classifier for a pixel of an input image x is denoted as $q^{(i)}(x) \in \mathbb{R}^C$, where C is the number of classes. The uncertainty of the distribution can be obtained as follows through the entropy equation.

$$H(i) = -\sum_{c=1}^{C} q_c^{(i)}(x) \log(q_c^{(i)}(x)). \tag{1}$$

The larger the entropy is, the more uncertain the distribution is. Thus, to reduce the effect of the uncertain distribution, we perform weighted-sum for the distribution, where the weight is inversely proportional to entropy by each distribution. Then, a new probability distribution $p^{(i)}(x) \in \mathbb{R}^C$ can be obtained in the form of a weighted sum of $q^{(i)}(x)$ as follows.

$$p(x) = \sum_{i=1}^{N} w_i \cdot q^{(i)}(x) \tag{2}$$

where w_i is the weight of i^{th} classifier, which is defined as

$$w_i = \frac{e^{-kH(i)}}{\sum_{i=1}^{N} e^{-kH(i)}}, \tag{3}$$

where k is a hyperparameter that controls the importance of the entropy

Then, the final class label c^* for each pixel is determined using the argmax operation as follows:

$$c^* = \arg\max_c p_c(x), \tag{4}$$

where $p_c(x)$ represents c^{th} class probability in $p(x)$. Finally, it is noteworthy that the model structure is not changed, so it can improve performance while using the existing pre-trained DDPM-based segmentation model.

IV. EXPERIMENTAL RESULTS

We use a total of 4 datasets including LSUN-Bedroom [11], FFHQ-256 [12], LSUN-Cat [11], and LSUN-Horse [11]. In the DDPM-based segmentation [6], 10 classifiers were respectively trained with freezing pre-trained ADM [10], and the number of images of the datasets used for the training was 40, 20, 30, and 30, respectively. Since our method proposes a new voting method, we compare it with the existing voting method used in the DDPM-based segmentation approach [6]. The overall pipeline for the DDPM-based segmentation remains the same, including the use of a pre-trained DDPM model. The only difference lies in the voting method employed. The performance evaluation of our method in terms of mean Intersection over Union (mIoU) is conducted on different datasets with varying numbers of test images. The datasets used for evaluation consist of 20 images for LSUN-Bedroom, 20 images for FFHQ-256, 20 images for LSUN-Cat, and 30 images for LSUN-Horse.

TABLE I
MEAN INTERSECTION OVER UNION(%) COMPARED TO PREVIOUS WORKS

Method	LSUN-Bedroom	FFHQ-256	LSUN-Cat	LSUN-Horse
DDPM-based [6]	50.2	57.8	57.5	64
Ours	**50.4**	**58**	**57.7**	**64.2**

Table I demonstrates the superiority of our proposed method over the existing method in terms of mean Intersection over Union (mIoU).

Fig. 2 shows comparison results with [6] and Our method, where our method achieves a 0.8% higher mIoU score than the DDPM-based method for that LSUN-Bedroom image, and a 2.2% higher mIoU score for that LSUN-Cat image.

In the case of the bedroom image, our method shows better similarity to the ground-truth specifically around the

chair and in the window frame compared to the DDPM-based method. This indicates that our method captures the details and boundaries of these regions more accurately.

Similarly, for the cat image, our method exhibits better similarity to the ground-truth, particularly around the cat's butt, compared to the DDPM-based method. This suggests that our method is able to capture the shape and contours of the cat's body more effectively.

V. Conclusion

We analyzed limitations and underlying factors in the ensemble process of previous DDPM-based method and proposed a new method to overcome the problem of the ensemble process. Using the proposed method, the issues of distribution uncertainty and non-smooth boundaries, which were problematic, are resolved. Experiments using the proposed method showed superior prediction performance than the previous work in terms of mIoU. However, this method cannot be considered to be a completely optimized model combination in the ensemble process, and there is much room for improvement in this regard.

Acknowledgment

This work has been supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2023-2018-0-01424) supervised by the IITP(Institute for Information communications Technology Promotion).

References

[1] Nontawat Tritrong, Pitchaporn Rewatbowornwong, and Supasorn Suwajanakorn. Repurposing gans for one-shot semantic part segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4475–4485, 2021.

[2] Jianjin Xu and Changxi Zheng. Linear semantics in generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9351–9360, 2021.

[3] Danil Galeev, Konstantin Sofiiuk, Danila Rukhovich, Mikhail Romanov, Olga Barinova, and Anton Konushin. Learning high-resolution domain-specific representations with a gan generator. In *Structural, Syntactic, and Statistical Pattern Recognition: Joint IAPR International Workshops, S+ SSPR 2020, Padua, Italy, January 21–22, 2021, Proceedings*, pages 108–118. Springer, 2021.

[4] Yuxuan Zhang, Huan Ling, Jun Gao, Kangxue Yin, Jean-Francois Lafleche, Adela Barriuso, Antonio Torralba, and Sanja Fidler. Datasetgan: Efficient labeled data factory with minimal human effort. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10145–10155, 2021.

[5] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[6] Dmitry Baranchuk, Ivan Rubachev, Andrey Voynov, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. *arXiv preprint arXiv:2112.03126*, 2021.

[7] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015.

[8] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

[9] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[10] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

[11] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

[12] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.