Annotator Consensus Prediction for Medical Image Segmentation with Diffusion Models

Tomer Amit, Shmuel Shichrur, Tal Shaharabany(✉), and Lior Wolf

Tel-Aviv University, Tel Aviv, Israel
{tomeramit1,shmuels1,shaharabany,wolf}@mail.tau.ac.il

Abstract. A major challenge in the segmentation of medical images is the large inter- and intra-observer variability in annotations provided by multiple experts. To address this challenge, we propose a novel method for multi-expert prediction using diffusion models. Our method leverages the diffusion-based approach to incorporate information from multiple annotations and fuse it into a unified segmentation map that reflects the consensus of multiple experts. We evaluate the performance of our method on several datasets of medical segmentation annotated by multiple experts and compare it with the state-of-the-art methods. Our results demonstrate the effectiveness and robustness of the proposed method. Our code is publicly available at https://github.com/tomeramit/Annotator-Consensus-Prediction.

Keywords: Multi annotator · Image segmentation · Diffusion Model

1 Introduction

Medical image segmentation is a challenging task that requires accurate delineation of structures and regions of interest in complex and noisy images. Multiple expert annotators are often employed to address this challenge, to provide binary segmentation annotations for the same image. However, due to differences in experience, expertise, and subjective judgments, annotations can vary significantly, leading to inter- and intra-observer variability. In addition, manual annotation is a time-consuming and costly process, which limits the scalability and applicability of segmentation methods.

To overcome these limitations, automated methods for multi-annotator prediction have been proposed, which aim to fuse the annotations from multiple annotators and generate an accurate and consistent segmentation result. Existing approaches for multi-annotator prediction include majority voting [7], label fusion [3], and label sampling [12].

In recent years, diffusion models have emerged as a promising approach for image segmentation, for example by using learned semantic features [2]. By modeling the diffusion of image intensity values over the iterations, diffusion models

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14223, pp. 544–554, 2023.
https://doi.org/10.1007/978-3-031-43901-8_52

capture the underlying structure and texture of the images and can separate regions of interest from the background. Moreover, diffusion models can handle noise and image artifacts, and adapt to different image modalities.

In this work, we propose a novel method for multi-annotator prediction, using diffusion models for medical segmentation. The goal is to fuse multiple annotations of the same image from different annotators and obtain a more accurate and reliable segmentation result. In practice, we leverage the diffusion-based approach to create one map for each level of consensus. To obtain the final prediction, we average the obtained maps and obtain one soft map.

We evaluate the performance of the proposed method on a dataset of medical images annotated by multiple annotators. Our results demonstrate the effectiveness and robustness of the proposed method in handling inter- and intra-observer variability and achieving higher segmentation accuracy than the state-of-the-art methods. The proposed method could improve the efficiency and quality of medical image segmentation and facilitate the clinical decision-making process.

2 Related Work

Multi-annotator Strategies. Research attention has recently been directed towards the issues of multi-annotator labels [7,12]. During training, Jensen et al. [12] randomly sampled different labels per image. This method produced a more calibrated model. Guan et al. [7] predicted the gradings of each annotator individually and acquired the corresponding weights for the final prediction. Kohl et al. [15] used the same sampling strategy to train a probabilistic model, based on a U-Net combined with a conditional variational autoencoder. Another recent probabilistic approach [20] combines a diffusion model with KL divergence to capture the variability between the different annotators. In our work, we use consensus maps as the ground truth and compare them to other strategies.

Diffusion Probabilistic Models (DPM). [23] are a class of generative models based on a Markov chain, which can transform a simple distribution (e.g. Gaussian) to data sampled from a complex distribution. Diffusion models are capable of generating high-quality images that can compete with and even outperform the latest GAN methods [5,9,19,23]. A variational framework for the likelihood estimation of diffusion models was introduced by Huang et al. [11]. Subsequently, Kingma et al. [14] proposed a Variational Diffusion Model that produces state-of-the-art results in likelihood estimation for image density.

Conditional Diffusion Probabilistic Models. In our work, we use diffusion models to solve the image segmentation problem as conditional generation, given the image. Conditional generation with diffusion models includes methods for class-conditioned generation, which is obtained by adding a class embedding to the timestep embedding [19]. In [4], a method for guiding the generative process in DDPM is present. This method allows the generation of images based on a given reference image without any additional learning. In the domain

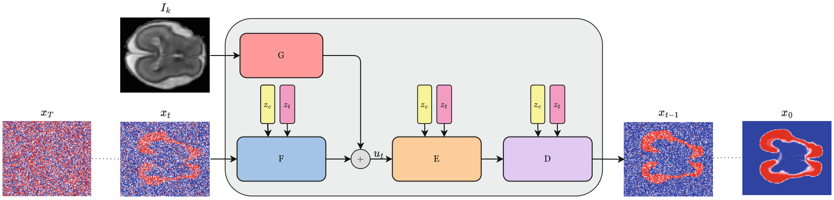

Fig. 1. The figure below illustrates our proposed method for multi-annotator segmentation. The input I_k image with the noisy segmentation map x_t is passed through our network iteratively T times in order to obtain an output segmentation map x_0. Each network receives the consensus level c as an embedding z_c as well as the time step data.

of super-resolution, the lower-resolution image is upsampled and then concatenated, channelwise, to the generated image at each iteration [10,21]. A similar approach passes the low-resolution images through a convolutional block [16] prior to the concatenation.

A previous study directly applied a diffusion model to generate a segmentation mask based on a conditioned input image [1]. Baranchuk et al. [2] extract features from a pretrained diffusion model for training a segmentation network, while our diffusion model generates the output mask. Compared to the diffusion-based image segmentation method of Wolleb et al. [26], our architecture differs in two main aspects: (i) the concatenation method of the condition signal, and (ii) an encoder that processes the conditioning signal. We also use a lower value of T, which reduces the running time.

3 Method

Our approach for binary segmentation with multi-annotators employs a diffusion model that is conditioned on the input image $I \in R^{W \times H}$, the step estimation t, and the consensus index c. The diffusion model updates its current estimate x_t iteratively, using the step estimation function ϵ_θ. See Fig. 1 for an illustration.

Given a set of C annotations $\{A_k^i\}_{i=1}^C$ associated with input sample I_k, we define the ground truth consensus map at level c to be

$$M_k^c[x,y] = \begin{cases} 1 & \sum_{i=1}^C A_k^i[x,y] \geq c, \\ 0 & \text{otherwise}, \end{cases} \tag{1}$$

During training, our algorithm iteratively samples a random level of the consensus $c \sim U[1,2,...,C]$ and an input-output pair (I_k, M_k^c). The iteration number $1 \leq t \leq T$ is sampled from a uniform distribution and X_T is sampled from a normal distribution.

We then compute x_t from X_T, M_k^c and t according to:

$$x_t = \sqrt{\bar{\alpha}_t} M_k^c + \sqrt{(1-\bar{\alpha}_t)} X_T, X_T \sim N(0, I_{n \times n}). \tag{2}$$

Algorithm 1. Training Algorithm

```
Input $T$, $D = \{(I_k, M_k^1, ..., M_k^C)\}_k^K$
repeat
    sample $c \sim \{1, ..., C\}$
    sample $(I_k, M_k^c) \sim D'$
    sample $\epsilon \sim N(\mathbf{0}, \mathbf{I_{n \times n}})$
    sample $t \sim (\{1,...,\mathrm{T}\})$
    $z_c = LUT_c(c)$
    $z_t = LUT_t(t)$
    $\beta_t = \frac{10^{-4}(T-t)+2*10^{-2}(t-1)}{T-1}$
    $\alpha_t = 1 - \beta_t$
    $\bar{\alpha}_t = \prod_{s=0}^{t} \alpha_s$
    $x_t = \sqrt{\bar{\alpha}_t} M_k^c + \sqrt{1 - \bar{\alpha}_t}\epsilon$
    $\nabla_\theta ||\epsilon - \epsilon_\theta(x_t, I_k, z_t, z_c)||$
until convergence
```

Algorithm 2. Inference Algorithm

```
Input $T$, $I$
for $c = 1, ..., C$ do
    sample $x_{T_c} \sim N(\mathbf{0}, \mathbf{I_{n \times n}})$
    for $t = T, T-1, ..., 1$ do
        sample $z \sim N(\mathbf{0}, \mathbf{I_{n \times n}})$
        $z_c = LUT_c(c)$, $z_t = LUT_t(t)$
        $\beta_t = \frac{10^{-4}(T-t)+2*10^{-2}(t-1)}{T-1}$
        $\alpha_t = 1 - \beta_t$. $\bar{\alpha}_t = \prod_{s=0}^{t} \alpha_s$
        $\tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$
        $\epsilon'_t = \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\bar{x}_t, I, z_t, z_c)$
        $\bar{x}_{t-1_c} = {\alpha_t}^{-\frac{1}{2}}(x_t - \epsilon'_t)$
        $x_{t-1_c} = \bar{x}_{t-1_c} + \mathbb{1}_{[t>1]}\tilde{\beta}_t^{\frac{1}{2}} z$
return $(\sum_{i=1}^{C} x_{0_i})/C$
```

where $\bar{\alpha}$ is a constant that defines the schedule of added noise.

The current step index t, and the consensus index c are integers that are translated to $z_t \in R^d$ and $z_c \in R^d$, respectively with a pair of lookup tables. The embeddings are passed to the different networks F, D and E.

In the next step, our algorithm encodes the input signal x_t with network F and encodes the condition image I_k with network G. We compute the conditioned signal $u_t = F(x_t, z_c, z_t) + G(I_k)$, and apply it to the networks E and D, where the output is the estimation of x_{t-1}.

$$\epsilon_\theta(x_t, I_k, z_t, z_c) = D(E(F(x_t, z_t, z_c) + G(I_k), z_t, z_c), z_t, z_c)\,. \tag{3}$$

The loss function being minimized is:

$$E_{x_0,\epsilon,x_e,t,c}[||\epsilon - \epsilon_\theta(x_t, I_k, z_t, z_c)||^2]. \tag{4}$$

The training procedure is depicted in Algorithm 1. The total number of diffusion steps T is set by the user, and C is the number of different annotators in the dataset. Our model is trained using binary consensus maps (M_k^c) as the ground truth, where k is the sample id, and c is the consensus index.

The inference process is described in Algorithm 2. We sample our model for each consensus index, and then calculate the mean of all results to obtain our target, which is a soft-label map representing the annotator agreement. Mathematically, if the consensus maps are perfect, this is equivalent to assigning each image location with the fraction of annotations that consider this location to be part of the mask (if c annotators mark a pixel, it would appear in levels $1..c$). In Sect. 4, we compare our method with other variants and show that estimating the fraction map directly, using an identical diffusion model, is far inferior to estimating each consensus level separately and then averaging.

Employing Multiple Generations. Since calculating x_{t-1} during inference includes the addition of $\mathbb{1}_{[t>1]}\tilde{\beta}_t^{\frac{1}{2}} z$ where z is from a standard distribution, there is significant variability between different runs of the inference method on the same inputs, see Fig. 2(b).

In order to exploit this phenomenon, we run the inference algorithm multiple times, then average the results. This way, we stabilize the results of segmentation and improve performance, as demonstrated in Fig. 2(c). We use twenty-five generated instances in all experiments. In the ablation study, we quantify the gain of this averaging procedure.

Architecture. In this architecture, the U-Net's decoder D is conventional and its encoder is broken down into three networks: E, F, and G. The last encodes the input image, while F encodes the segmentation map of the current step x_t. The two processed inputs have the same spatial dimensionality and number of channels. Based on the success of residual connections [8], we sum these signals $F(x_t, z_t, z_c) + G(I)$. This sum then passes to the rest of the U-Net encoder E.

The input image encoder G is built from Residual in Residual Dense Blocks [24] (RRDBs), which combine multi-level residual connections without batch normalization layers. G has an input 2D-convolutional layer, an RRDB with a residual connection around it, followed by another 2D-convolutional layer, leaky RELU activation and a final 2D-convolutional output layer. F is a 2D-convolutional layer with a single-channel input and an output of L channels.

The encoder-decoder part of ϵ_θ, i.e., D and E, is based on U-Net, similarly to [19]. Each level is composed of residual blocks, and at resolution 16×16 and 8×8 each residual block is followed by an attention layer. The bottleneck contains two residual blocks with an attention layer in between. Each attention layer contains multiple attention heads.

The residual block is composed of two convolutional blocks, where each convolutional block contains group-norm, SiLU activation, and a 2D-convolutional layer. The residual block receives the time embedding through a linear layer, SiLU activation, and another linear layer. The result is then added to the output of the first 2D-convolutional block. Additionally, the residual block has a residual connection that passes all its content.

On the encoder side (network E), there is a downsample block after the residual blocks of the same depth, which is a 2D-convolutional layer with a stride of two. On the decoder side (network D), there is an upsample block after the residual blocks of the same depth, which is composed of the nearest interpolation that doubles the spatial size, followed by a 2D-convolutional layer. Each layer in the encoder has a skip connection to the decoder side.

4 Experiments

We conducted a series of experiments to evaluate the performance of our proposed method for multi-annotator prediction. Our experiments were carried out on datasets of the QUBIQ benchmark[1]. We compared the performance of our proposed method with several state-of-the-art methods.

[1] Quantification of Uncertainties in Biomedical Image Quantification Challenge in MICCAI20'- link.

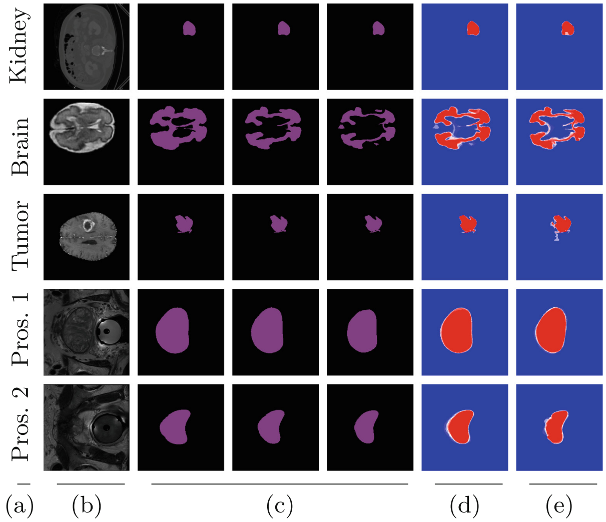

Fig. 2. Multiple segmentation results on all datasets of the QUBIQ benchmark. (a) dataset, (b) input image, (c) a subset of the obtained consensus maps for multiple runs with different consensus index on the same input, (d) average result, visualized by the 'bwr' color scale between 0 (blue) and 1 (red), and (e) ground truth. (Color figure online)

Datasets. The Quantification of Uncertainties in Biomedical Image Quantification Challenge (QUBIQ), is a recently available challenge dataset specifically for the evaluation of inter-rater variability. QUBIQ comprises four different segmentation datasets with CT and MRI modalities, including brain growth (one task, MRI, seven raters, 34 cases for training and 5 cases for testing), brain tumor (one task, MRI, three raters, 28 cases for training and 4 cases for testing), prostate (two subtasks, MRI, six raters, 33 cases for training and 15 cases for testing), and kidney (one task, CT, three raters, 20 cases for training and 4 cases for testing). Following [13], the evaluation is performed using the soft Dice coefficient with five threshold levels, set as (0.1, 0.3, 0.5, 0.7, 0.9).

Implementation Details. The number of diffusion steps in previous works was 1000 [9] and even 4000 [19]. The literature suggests that more is better [22]. In our experiments, we employ 100 diffusion steps, to reduce inference time.

The AdamW [18] optimizer is used in all our experiments. Based on the intuition that the more RRDB blocks, the better the results, we used as many blocks as we could fit on the GPU without overly reducing batch size.

Following [13], for all datasets of the QUBIQ benchmark the input image resolution, as well as the test image resolution, was 256 × 256. The experiments were performed with a batch size of four images and eight RRDB blocks. The network depth was seven, and the number of channels in each depth was $[L, L, L, 2L, 2L, 4L, 4L]$, with $L = 128$. The augmentations used were: random

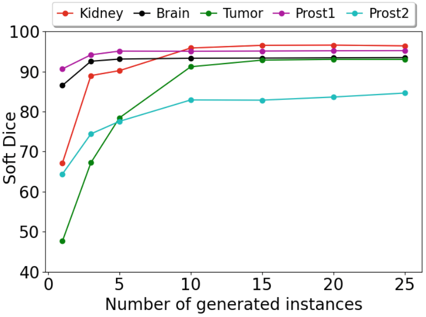

Fig. 3. Soft Dice vs. #generated images.

Table 1. QUBIQ soft Dice results.

Method	Kidney	Brain	Tumor	Prost1	Prost2
FCN	70.03	80.99	83.12	84.55	67.81
MCD	72.93	82.91	86.17	86.40	70.95
FPM	72.17	–	–	–	–
DAF	–	–	–	85.98	72.87
MV-UNet	70.65	81.77	84.03	85.18	68.39
LS-UNet	72.31	82.79	85.85	86.23	69.05
MH-UNet	73.44	83.54	86.74	87.03	75.61
MRNet	74.97	84.31	88.40	87.27	76.01
AMIS	68.53	74.09	92.95	91.64	21.91
DMISE	74.50	92.80	87.80	94.70	80.20
Ours	**96.58**	**93.81**	**93.16**	**95.21**	**84.62**

scaling by a factor sampled uniformly in the range $[0.9, 1.1]$, a rotation between 0 and 15°, translation between $[0, 0.1]$ in both axes, and horizontal and vertical flips, each applied with a probability of 0.5.

Results. We compare our method with FCN [17], MCD [6], FPM [27], DAF [25], MV-UNet [13], LS-UNet [12], MH-UNet [7], and MRNet [13].

We also compare with models that we train ourselves, using public code AMIS [20], and DMISE [26]. The first is trained in a scenario where each annotator is a different sample ("No annotator" variant of our ablation results below), and the second is trained on the consensus setting, similar to our method. As can be seen in Table 1, our method outperforms all other methods across all datasets of QUBIQ benchmark.

Ablation Study. We evaluate alternative training variants as an ablation study in Table 2. The "Annotator" variant, in which our model learns to produce each annotator binary segmentation map and then averages all the results to obtain the required soft-label map, achieves lower scores compared to the "Consensus" variant, which is our full method. The "No annotator" variant, where images were paired with random annotators without utilizing the annotator IDs, achieves a slightly lower average score compared to the "Annotator" variant. We also note that our "No annotator" variant outperforms the analog AMIS model in four out of five datasets, indicating that our architecture is somewhat preferable. In a third variant, our model learns to predict the soft-label map that denotes the fraction of annotators that mark each image location directly. Since this results in fewer generated images, we generate C times as many images per test sample. The score of this variant is also much lower than that of our method.

Next, we study the effect of the number of generated images on performance. The results can be seen in Fig. 3. In general, increasing the number of generated instances tends to improve performance. However, the number of runs required to reach optimal performance varies between classes. For example, for the Brain and the Prostate 1 datasets, optimal performance is achieved using 5 generated images, while on Prostate 2 the optimal performance is achieved using 25 gen-

Table 2. Ablation study showing soft Dice results for various alternative methods of training similar diffusion models.

Method	Kidney	Brain	Tumor	Prostate 1	Prostate 2
Annotator	96.13	89.88	92.51	93.89	76.89
No annotator	94.46	89.78	91.78	92.58	78.61
Soft-label	65.41	79.56	75.60	73.23	65.24
Consensus (our method)	**96.58**	**93.81**	**93.16**	**95.21**	**84.62**

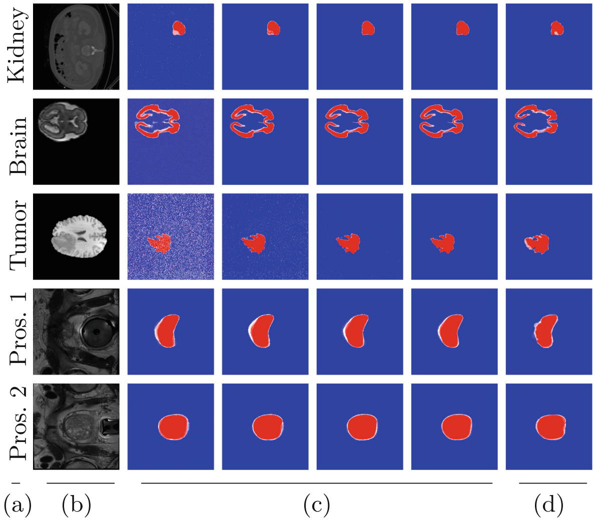

Fig. 4. Multiple segmentation results per number of generated images. (a) dataset, (b) input image, (c) results for 1, 5, 10, 25 generated images, and (d) ground truth.

erated images. Figure 4 depicts samples from multiple datasets and presents the progression as the number of generated images increases. As can be seen, as the number of generated images increases, the outcome becomes more and more similar to the target segmentation.

5 Discussion

In order to investigate the relationship between the annotator agreement and the performance of our model, we conducted an analysis by calculating the average Dice score between each pair of annotators across the entire dataset. The results of this pairwise Dice analysis can be found in Table 3, where higher mean-scores indicate a greater consensus among the annotators.

We observed that our proposed method demonstrated improved performance on datasets with higher agreement among annotators, specifically the kidney and

Table 3. Pairwise Dice scores per dataset.

Dataset	Mean score between pairs
Kidney	94.95
Brain	85.74
Tumor	90.65
Prostate 1	94.64
Prostate 2	89.91

prostate 1 datasets. Conversely, the performance of the other methods significantly deteriorated on the kidney dataset, leading to a lower correlation between the Dice score and the overall performance.

Additionally, we examined the relationship between the number of annotators and the performance of our model. Surprisingly, we found no significant correlation between the number of annotators and the performance of our model.

6 Conclusions

Shifting the level of consensus required to mark a region from very high to as low as one annotator, can be seen as creating a dynamic shift from a very conservative segmentation mask to a very liberal one. As it turns out, this dynamic is well-captured by diffusion models, which can be readily conditioned on the level of consensus. Another interesting observation that we make is that the mean (over the consensus level) of the obtained consensus masks is an effective soft mask. Applying these two elements together, we obtain state-of-the-art results on multiple binary segmentation tasks.

Acknowledgment. This project has received funding from the ISRAEL SCIENCE FOUNDATION (grant No. 2923/20) within the Israel Precision Medicine Partnership program. It was also supported by a grant from the Tel Aviv University Center for AI and Data Science (TAD).

References

1. Amit, T., Nachmani, E., Shaharbany, T., Wolf, L.: SegDiff: image segmentation with diffusion probabilistic models. arXiv preprint arXiv:2112.00390 (2021)
2. Baranchuk, D., Rubachev, I., Voynov, A., Khrulkov, V., Babenko, A.: Label-efficient semantic segmentation with diffusion models. arXiv preprint arXiv:2112.03126 (2021)
3. Chen, G., et al.: Automatic pathological lung segmentation in low-dose CT image using eigenspace sparse shape composition. IEEE Trans. Med. Imaging **38**(7), 1736–1749 (2019)
4. Choi, J., Kim, S., Jeong, Y., Gwon, Y., Yoon, S.: ILVR: conditioning method for denoising diffusion probabilistic models. arXiv preprint arXiv:2108.02938 (2021)

5. Dhariwal, P., Nichol, A.: Diffusion models beat GANs on image synthesis. In: Advances in Neural Information Processing Systems, vol. 34 (2021)
6. Gal, Y., Ghahramani, Z.: Dropout as a Bayesian approximation: representing model uncertainty in deep learning. In: International Conference on Machine Learning, pp. 1050–1059. PMLR (2016)
7. Guan, M., Gulshan, V., Dai, A., Hinton, G.: Who said what: modeling individual labelers improves classification. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32 (2018)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
9. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Adv. Neural. Inf. Process. Syst. **33**, 6840–6851 (2020)
10. Ho, J., Saharia, C., Chan, W., Fleet, D.J., Norouzi, M., Salimans, T.: Cascaded diffusion models for high fidelity image generation. J. Mach. Learn. Res. **23**(47), 1–33 (2022)
11. Huang, C.W., Lim, J.H., Courville, A.C.: A variational perspective on diffusion-based generative models and score matching. In: Advances in Neural Information Processing Systems, vol. 34 (2021)
12. Jensen, M.H., Jørgensen, D.R., Jalaboi, R., Hansen, M.E., Olsen, M.A.: Improving uncertainty estimation in convolutional neural networks using inter-rater agreement. In: Shen, D., et al. (eds.) MICCAI 2019, Part IV. LNCS, vol. 11767, pp. 540–548. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32251-9_59
13. Ji, W., et al.: Learning calibrated medical image segmentation via multi-rater agreement modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12341–12351 (2021)
14. Kingma, D.P., Salimans, T., Poole, B., Ho, J.: Variational diffusion models. arXiv preprint arXiv:2107.00630 (2021)
15. Kohl, S., et al.: A probabilistic U-Net for segmentation of ambiguous images. In: Advances in Neural Information Processing Systems, vol. 31 (2018)
16. Li, H., et al.: SRDIFF: single image super-resolution with diffusion probabilistic models. Neurocomputing **479**, 47–59 (2022)
17. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440 (2015)
18. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
19. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: International Conference on Machine Learning, pp. 8162–8171. PMLR (2021)
20. 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, pp. 11536–11546 (2023)
21. Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D.J., Norouzi, M.: Image super-resolution via iterative refinement. arXiv preprint arXiv:2104.07636 (2021)
22. San-Roman, R., Nachmani, E., Wolf, L.: Noise estimation for generative diffusion models. arXiv preprint arXiv:2104.02600 (2021)
23. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International Conference on Machine Learning, pp. 2256–2265. PMLR (2015)

24. Wang, X., et al.: ESRGAN: enhanced super-resolution generative adversarial networks. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops, pp. 0–0 (2018)
25. Wang, Y., et al.: Deep attentional features for prostate segmentation in ultrasound. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018, Part IV. LNCS, vol. 11073, pp. 523–530. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00937-3_60
26. Wolleb, J., Sandkühler, R., Bieder, F., Valmaggia, P., Cattin, P.C.: Diffusion models for implicit image segmentation ensembles (2021)
27. Zhou, Y., Xie, L., Shen, W., Wang, Y., Fishman, E.K., Yuille, A.L.: A fixed-point model for pancreas segmentation in abdominal CT scans. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI 2017, Part I. LNCS, vol. 10433, pp. 693–701. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66182-7_79