Enhancing Label-efficient Medical Image Segmentation with Text-guided Diffusion Models

Chun-Mei Feng[1]

[1] Institute of High Performance Computing (IHPC),
Agency for Science, Technology and Research (A*STAR)
fengcm.ai@gmail.com
https://github.com/chunmeifeng/TextDiff

Abstract. Aside from offering state-of-the-art performance in medical image generation, denoising diffusion probabilistic models (DPM) can also serve as a representation learner to capture semantic information and potentially be used as an image representation for downstream tasks, *e.g.,* segmentation. However, these latent semantic representations rely heavily on labor-intensive pixel-level annotations as supervision, limiting the usability of DPM in medical image segmentation. To address this limitation, we propose an enhanced diffusion segmentation model, called TextDiff, that improves semantic representation through inexpensive medical text annotations, thereby explicitly establishing semantic representation and language correspondence for diffusion models. Concretely, TextDiff extracts intermediate activations of the Markov step of the reverse diffusion process in a pretrained diffusion model on large-scale natural images and learns additional expert knowledge by combining them with complementary and readily available diagnostic text information. TextDiff freezes the dual-branch multi-modal structure and mines the latent alignment of semantic features in diffusion models with diagnostic descriptions by only training the cross-attention mechanism and pixel classifier, making it possible to enhance semantic representation with inexpensive text. Extensive experiments on public QaTa-COVID19 and MoNuSeg datasets show that our TextDiff is significantly superior to the state-of-the-art multi-modal segmentation methods with only a few training samples.

Keywords: Medical image segmentation · Diffusion model · Language and image

1 Introduction

The denoising Diffusion Probability Model (DPM) has recently demonstrated state-of-the-art performance in medical image generation [11,14,22,9,5,14], *e.g.,* synthesis of pathological images [19], generation of 3D brain MRI [20,**?**], and dynamic disease progression fitting [15], even surpassing GAN-based approaches.

Interestingly, recent work has found the potential of DPM as a representation learner to capture semantic information, as well as its advantages for downstream tasks such as natural image segmentation [2]. However, obtaining high-quality medical images necessary for lesion segmentation is difficult, and their pixel-level labeling is labor-intensive. As a result, the performance of deep learning-based medical image segmentation models, including DPM, is significantly limited [23]. This trend highlights the bottleneck caused by an over-reliance on labor-intensive pixel-level annotations as supervision to mine latent semantic representations. Instead, techniques like semi-supervised learning [24,17] and weakly supervised learning [10] are being applied to reduce the deep model's dependence on large amounts of annotated data. Unfortunately, the effectiveness of these techniques heavily relies on the confidence of the pseudo-labels. If a large number of pseudo-labels have low confidence, the segmentation accuracy can be significantly hampered [18], which greatly limits the clinical applicability of deep learning techniques. *Therefore, how to develop an effective label-efficient diffusion model for medical image segmentation remains an unresolved question.*

As a remedy, we seek to increase the data usability by extracting knowledge from other readily available sources of medical information, such as text diagnostic information, to complement medical images. Medical text records are usually generated alongside sampled images, and accessing text diagnostic information corresponding to the images incurs no additional cost [18]. The text diagnostic information records additional information complementary to image data. Huang *et al.* leverage the radiology reports to learn global and local representations by contrasting image sub-regions and text annotations [12], while Li *et al.* introduce the medical text annotation to compensate the vision transformer [18]. These methods demonstrate the usefulness of text diagnosis in image diagnosis using deep learning technology. *Despite recent progress, it is unclear whether medical text diagnosis can also benefit the performance of diffusion models on medical image segmentation.* Hence, we investigate how additional medical text diagnostic information can directly address the aforementioned issues by incorporating it into the diffusion model [12].

In this paper, we improve the performance of diffusion models in medical segmentation from the perspective of mining inexpensive medical text diagnostic information, yielding a new algorithm TextDiff that exhibits strong performance compared to various state-of-the-art multi-modal segmentation algorithms. Our main contributions are as follows:

1. We propose an enhanced label-efficient medical image segmentation method, termed TextDiff, to reduce the dependence of the diffusion model on pixel-level annotations by learning additional expert knowledge through medical text annotations.
2. We establish strong connections between *textual diagnostic annotations* and *intermediate activations* of the Markov step of the reverse diffusion process in DPM, thereby improving visual-semantic representations in diffusion models.
3. We *freeze* the two-branch structure of TextDiff while *only training* the cross-attention and pixel classifier, yielding significantly better results than various state-of-the-art multi-modal segmentation methods on COVID and

pathological images with very few training samples, *e.g.,* compared with GLoRIA [12], TextDiff obtain the results of Dice: 66.38% → **78.67%** and IoU: 49.83% → **64.98%** on MoNuSeg dataset.

2 Methodology

2.1 Overall Architecture

Given a medical image to be segmented, our goal is to train a deep neural network to automatically localize the visual region spatially of a certain tissue or lesion that the doctors are interested in. Here, unlike the previous works on diffusion models in image generation, we further explore the ability of DPM to capture high-level semantic information. Existing works train the network using visual information [2]; on the contrary, we enhance the visual-semantic information by introducing inexpensive text diagnostic annotations that provide more efficient results. Such mechanism reduces the segmentation model's reliance on pixel-level annotations.

Since these texts are generated simultaneously with the diagnostic images, our training samples can be expressed as $\mathcal{D} = \left\{ \left(\mathbf{x}^1, \mathbf{t}^1\right), \left(\mathbf{x}^2, \mathbf{t}^2\right), \ldots, \left(\mathbf{x}^N, \mathbf{t}^N\right) \right\}$, where $\mathbf{x}$, $\mathbf{t}$ refer to the diagnostic images (*e.g.,* CT, X-Ray, or MRI images) and their corresponding text diagnostic annotation, respectively. As shown in Fig. 1, the proposed TextDiff extracts visual and textual information by vision-language dual-branch architecture, *i.e.,* diffusion model with UNet architecture [7] and `Clinical BioBERT` [1], respectively, and finally establish connections between textual diagnostic information and intermediate activations of the Markov step of the reverse diffusion process in DPM. Specifically, our TextDiff accepts two different modalities as input, *i.e.,* $\mathbf{x}$ and $\mathbf{t}$. Each modality is sent to the pre-trained model to obtain the multi-modal feature representations, which are fused and then used to obtain predicted segmentation results by the pixel classifier. Note that only the multi-scale cross-attention and the pixel classifier are trainable in our method. We fix the weights of both the text encoder and the image encoder to maintain the vision-language alignment. To demonstrate the effectiveness of our proposed method, here, we simply use the means dice loss $\mathcal{L}_{\texttt{Dice}}$ and cross-entropy loss $\mathcal{L}_{\texttt{CE}}$ to evaluate the segmentation performance.

Image Encoding. For a scanned image $\mathbf{x}$, we can obtain the features by a visual backbone $\hat{\mathbf{x}} = \mathcal{E}_{\texttt{Im}}(\mathbf{x})$. In our method, we explore whether the diffusion models can serve as an powerful instrument for segmentation. In image generation, diffusion models are used to transform noise $\mathbf{x}_T \sim N(0, I)$ to the sample $\mathbf{x}_0$ by gradually denoising $\mathbf{x}_T$ to less noisy samples $\mathbf{x}_t$. Formally, the forward diffusion process can be expressed as:

$$q\left(\mathbf{x}_t \mid \mathbf{x}_{t-1}\right) := \mathcal{N}\left(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t I\right). \tag{1}$$

where $\beta_1, \ldots, \beta_t$ are the fixed variance schedule.

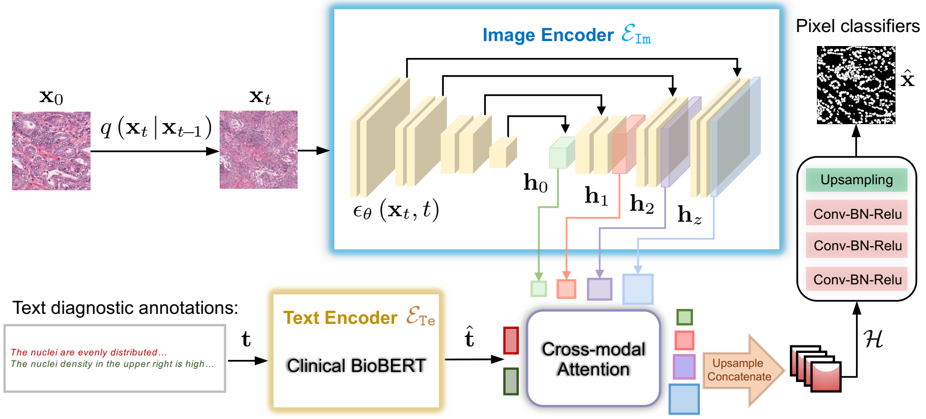
Image Encoder $\mathcal{E}_{\mathtt{Im}}$
Pixel classifiers
$\hat{\mathbf{x}}$
$\mathbf{x}_0$
$q\left(\mathbf{x}_t \mid \mathbf{x}_{t-1}\right)$
$\mathbf{x}_t$
$\epsilon_\theta\left(\mathbf{x}_t, t\right)$
$\mathbf{h}_0$
$\mathbf{h}_1$
$\mathbf{h}_2$
$\mathbf{h}_z$
Upsampling
Conv-BN-Relu
Conv-BN-Relu
Conv-BN-Relu
Text diagnostic annotations:
The nuclei are evenly distributed…
The nuclei density in the upper right is high…
$\mathbf{t}$
Text Encoder $\mathcal{E}_{\mathtt{Te}}$
Clinical BioBERT
$\hat{\mathbf{t}}$
Cross-modal Attention
Upsample Concatenate
$\mathcal{H}$

Fig. 1. Overview of the proposed TexDiff framework, where the *Image Encoder* is based on a pre-trained `Diffusion` [7] model to produce the high-level semantic information, while `Clinical BioBERT` [1] serves as the *Text Encoder*. *Multi-scale Cross-modal Attention* leverages the knowledge of the text diagnostic annotation and images to be aligned for enhancing semantic representations.

Mathematically, the pre-trained DPM approximates a reverse process which can be expressed as follows

$$p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right):=\mathcal{N}\left(\mathbf{x}_{t-1} ; \mu_\theta\left(\mathbf{x}_t, t\right), \Sigma_\theta\left(\mathbf{x}_t, t\right)\right). \tag{2}$$

Here, for an image input $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$, we can compute T sets of activation tensors from the noise predictor $\epsilon_\theta\left(\mathbf{x}_t, t\right)$ which is typically parameterized by different variants of the UNet architecture [7]. As shown in Fig. 1, we first add Gaussian noise to corrupt $\mathbf{x}_0$. The parameterization of the UNet model $\epsilon_\theta\left(\mathbf{x}_t, t\right)$ uses the noisy $\mathbf{x}_t$ as an input. Bilinear interpolation is then used to upsample the intermediate activations of the UNet to $H \times W$. Such mechanism enables them to be treated as pixel-level representations of $\mathbf{x}_0$.

Based on this, we further extract the pixel-level representation of the labeled image through UNet blocks and diffusion steps t, *e.g.*, the middle block $B = \{4, 6, 7, 8, 12, 16\}$ of UNet decoder and steps $t = \{50, 150, 250\}$ are adopt for the feature extraction [2]. Then, to create feature vectors $\hat{\mathbf{x}}$ for all of the pixels in the training images, the extracted representations from these blocks and steps are fused with medical text features $\hat{\mathbf{t}}$ produced by text encoder $\mathcal{E}_{\mathtt{Te}}$. Note that we only consider decoder activations because the skip connections also collect encoder activations. Finally, we can obtain the prediction results $\hat{\mathbf{x}}$ for each pixel by training the pixel classifier while freezing the two-branch backbone.

Text Encoding. As we mentioned before, the textual diagnostic annotations are generated without extra collection cost alongside the sampled images, and their small footprint makes them a natural complement to image data. As a result, as shown in Fig. 1, we use a pre-trained text encoder, *i.e.*, `Clinical BioBERT` [1], to extract valuable information from the text diagnostic annotations. `Clinical`

BioBERT [1] is the pre-trained text model which obtain the clinical-aware text embeddings on the MIMIC III dataset [13]. Specifically, given a annotations $\mathbf{t}$, we can obtain the features by a text backbone $\hat{\mathbf{t}} = \mathcal{E}_{\texttt{Te}}(\mathbf{t})$. Subsequently, the text features are fed into the cross-modal attention along with the intermediate activations of the Markov step of the reverse diffusion process in DPM. Details are provided in Sec. 2.2.

2.2 Cross-modal Attention for Knowledge Alignment

Here, we consider how to align the text and visual features, thereby enhancing the visual semantic representation. We define a cross-modal attention module $\mathcal{M}_{\texttt{cro}}$ that integrates features of different sizes in the diffusion model of an UNet shape, thereby cross-contextualizing the text embeddings with pixel representations of the image. The cross-modal attention module provides a strong connection between language and vision, enabling textual information to enhance semantic representation in images [8]. Specifically, given the pixel-level visual features, $\mathbf{h}_z$, where $i = 0, 1, ..., z$, with different sizes extracted from the diffusion decoder and its corresponding text feature $\hat{\mathbf{t}}$, we compute the scaled dot-product attention at the step t:

$$\mathcal{H}_{z,t} = \text{Softmax}\left(\mathbf{h}_{z,t} W_q (\hat{\mathbf{t}} W_k)^T / \sqrt{d}\right) \hat{\mathbf{t}} W_v, \tag{3}$$

where W_q, W_k, and W_v are the learned parameter matrices. $\mathcal{H}_{z,t}$ is the different scales attention representation over the text enhanced visual medical image. We concatenate these attention representations to get $\mathcal{H}$ after upsampling them to the same size. We analyze whether medical text diagnosis brings benefits to the performance of diffusion models on medical image segmentation in Sec. 3.2.

3 Experiments

Experimental Setup. We train our method by Pytorch with one NVIDIA Tesla V100 GPU and 32GB of memory. We use Adam as the optimizer, with an initial learning rate of 1e−4 and a batch size of 1, for 100 epochs. The input images are resized to 256 × 256. The middle blocks $B = \{6, 8, 12, 16\}$ and $B = \{4, 6, 8, 12\}$ of the UNet decoder with steps $t = \{50, 150, 250\}$ are adopted for MoNuSeg and QaTa-COVID19, respectively.

Datasets. We employ two public datasets to evaluate our method, *i.e.,* **1) MoNuSeg** [16] is a pathology dataset obtained from the MICCAI 2018 MoNuSeg challenge and consists of 30 images with 21,623 nuclear boundary annotations for training and 14 images with 7000 nuclear boundary annotations for testing. To demonstrate the effectiveness of our label-efficient segmentation mechanism, we experiment with only a small number of images to clearly demonstrate the advantages of our method. We randomly select 5 images from **MoNuSeg** for `training`, while the test set remains unchanged; **2) QaTa-COVID19** [6] is collected from Qatar University and Tampere University and consist of 9258 COVID-19 chest radiographs with pixel annotations of COVID-19 lesions. In

Table 1. Quantitative comparison of state-of-the-art methods on two datasets, where # Param is the parameter cost, ↑ and ↓ indicate increments and decrements compared with UNet, respectively. Detailed analyses are provided in Sec.3.1.

		MoNuSeg		QaTa-COVID19	
Method	# Param.	Dice (%)	IoU (%)	Dice (%)	IoU (%)
UNet$_{2015}$[21]	31.04 M	73.92(00.00)	58.98(00.00)	46.08(00.00)	34.16(00.00)
TransUNet$_{2021}$[4]	93.19 M	73.54(00.38) ↓	58.79(00.19) ↓	70.78(24.70) ↑	59.50(25.34) ↑
SwinUNet$_{2021}$[3]	27.17 M	64.36(09.56) ↓	48.74(10.24) ↓	65.19(19.11) ↑	51.87(17.71) ↑
GLoRIA$_{2021}$[12]	32.52 M	66.38(07.54) ↓	49.83(09.15) ↓	71.05(24.97) ↑	**59.74**(24.58) ↑
LViT$_{2022}$[18]	29.72 M	57.95(15.97) ↓	44.13(14.85) ↓	66.43(20.35) ↑	51.71(17.55) ↑
TextDiff (Ours)	**9.68** M	**78.67**(04.75) ↑	**64.98**(06.00) ↑	**71.41**(25.33) ↑	59.03(24.87) ↑

our experiments, we randomly select 150 images for `training`. Following [18], we use their extended text annotations to enhance the vision-language model.

Baselines. We compare two categories of methods to demonstrate the effectiveness of our proposed method, including **a)** classical medical segmentation methods: (1) UNet [21], (2) TransUNet [4], (3) SwinUNet[3], and **b)** text-driven medical segmentation methods: (1) GLoRIA [12], a multi-modal medical image recognition framework that learns the global and local representations by contrasting image sub-regions and text in the paired report; (2) LViT [18], a multi-modal medical image segmentation framework based on the transformer that uses the medical text annotations to compensate for the visual representation. For a fair comparison, we retrain all the baseline methods with their default parameter and report the best results.

3.1 Comparison with State-of-the-arts.

To investigate the effectiveness of our method, we show the comparison results, *i.e.,* Dice (%) and IoU (%), with various state-of-the-art methods in Table 1. Our approach yields the highest values on all datasets with regard to Dice (%) and IoU (%). Specifically, the classical medical segmentation methods, *i.e.,* UNet [21], and TransUNet [4], SwinUNet[3] are less effective than the language-vision methods, *i.e.,* GLoRIA [12], LViT [18], and our proposed method. However, the segmentation results of multi-modal methods, *i.e.,* GLoRIA [12] and LViT [18], are still lower than our method. This is mainly due to the deep alignment of text and visual information extracted by the diffusion model in our method. Although both the GLoRIA [12] and LViT [18] absorbed the text information, LViT requires more parameters, *i.e.,* 29.72 M, and GLoRIA cannot provide effective visual features. Besides, LViT needs to be trained from scratch, while our model takes advantage of the powerful large-scale pre-training model on natural images. The dual-branch text and image encoders of our method are frozen, and we only need to update the pixel classifier and cross-attention mechanism. In particular, compared with the state-of-the-art vision-language medical method

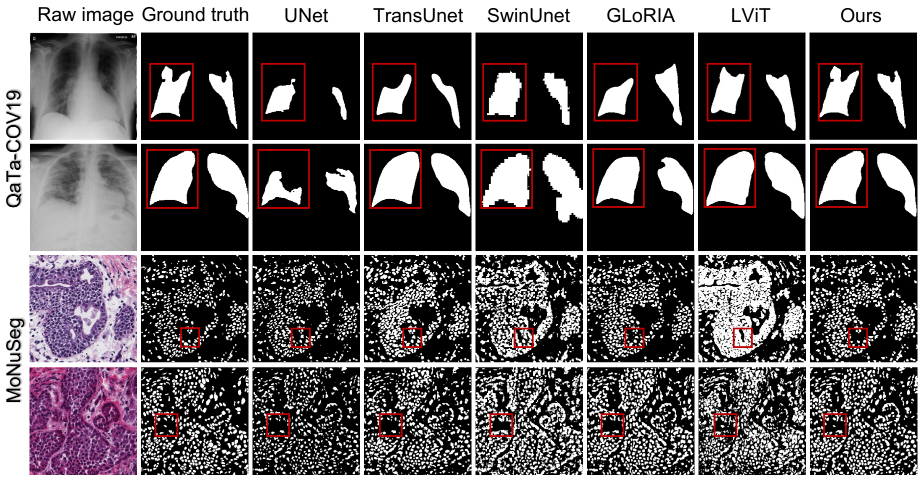

Fig. 2. Visual segmentation comparisons with regards to different datasets, detailed analysis is provided in Sec. 3.1.

GLoRIA [12], our method improves the Dice (%) values from 66.38 to **78.67**, and the IoU (%) values from 49.83 to **64.98** on the MONuSeg dataset. More importantly, our proposed method only requires 9.68 M of communication. These demonstrate that our model can effectively enhance the visual features through the text annotations, which is beneficial to medical image segmentation and reduces reliance on labor-intensive pixel-level annotation as supervision.

Fig. 2 provides the qualitative results of our model and other state-of-the-art methods on the two datasets. As compared with the ground-truth, the language-vision methods outperform the classical medical segmentation methods. Notably, our method provides fewer errors and segmentation results that are closest to ground-truth, which is attributable to the fact that our method can effectively learn enhanced features from the inexpensive text annotations. Further, these visual segmentation results from different datasets, *i.e.,* MoNuSeg [16] and QaTa-COVID19 [6], support our conclusion that our method can leverage the medical text diagnosis to benefit the performance of diffusion models on medical image segmentation.

3.2 Ablation Studies.

Representation analysis. To investigate the performance of medical text diagnostic annotations in the diffusion representations, we show the evolution of prediction performance over different blocks and diffusion steps t in Fig. 3, where (a) and (c) are the representations without text, (b) and (d) are the representations with text. As can be seen from this figure, the semantic representations produced by the diffusion model vary for different blocks and diffusion steps, and these representations are enhanced to varying degrees after introducing text diagnostic annotations (see Fig. 3 (b) and (c)). These experimental results provide

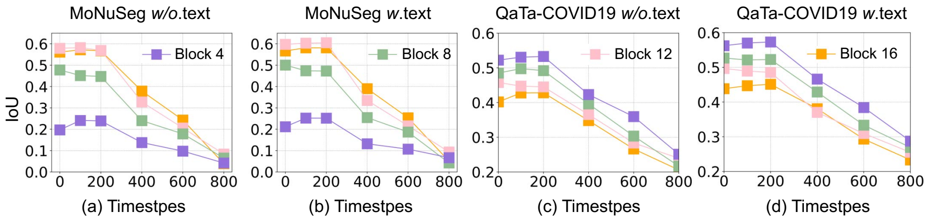

Fig. 3. Evolution of the segmentation performance with regard to different blocks and steps of our proposed method on the two datasets, see Sec.3.2 for more details.

Table 2. Ablation studies on the MoNuSeg dataset, where ↓ indicates decrements compared with our full model TextDiff. Detailed analyses are provided in Sec.3.2.

			MoNuSeg		QaTa-COVID19	
Variation	`Text`	$\mathcal{M}_{\texttt{cro}}$	Dice (%)	IoU (%)	Dice (%)	IoU (%)
ζ_1	-	-	77.73(0.73) ↓	63.70(0.95) ↓	68.90(5.37) ↓	56.94(5.09) ↓
ζ_2	✓	-	76.34(2.12) ↓	61.89(2.76) ↓	70.71(3.56) ↓	57.81(4.22) ↓
`Ours`	✓	✓	78.67(0.00)	64.98(0.00)	71.41(0.00)	59.03(0.00)

an ***answer*** to the question of whether medical text diagnosis can improve the performance of diffusion models on medical image segmentation. Additionally, we find that features corresponding to later steps in the reverse diffusion process are often more effective at capturing semantic information, while the ones corresponding to earlier steps are generally uninformative. In different blocks, the features produced by different layers of the UNet decoder on the two datasets seem to be different, allowing us to choose different blocks for different datasets in our experiments.

Key components analysis. Here, we evaluate the effectiveness of the key components of our method through some variations of it, *i.e.*, ζ_1, which is our method without medical text annotations, and ζ_2, which is our method without multi-scale cross attention. We summarize the results of these ablation models in Table 2. As can be seen from this table, we observe that ζ_1 performs the worst, which is consistent with our primary motivation that the text annotations can provide supplementary information for visual features. Since without multi-scale cross-attention module cannot learn the deep information of the two modalities, the results of ζ_2 are not optimal. Yet, our full TextDiff further aligns the features of the different modalities, and yields the best results, demonstrating its powerful capability in medical image segmentation.

4 Conclusion

This work focuses on how to extend the diffusion model to the medical segmentation task with text annotations, thereby reducing the over-reliance on labor-intensive pixel-level annotations as supervision. Given this, we propose a diffusion segmentation method, termed TextDiff, that improves semantic representation via inexpensive medical text annotations, allowing the model to perform well on a small number of training images. To the best of our knowledge, TextDiff is the first multi-modal diffusion framework for medical image segmentation. Experiments on the different datasets demonstrate the superiority of TextDiff in medical image segmentation with limited training samples.

Acknowledgement

This research/project is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG2-TC-2021-003). This work was supported by the Agency for Science, Technology and Research (A*STAR) through its AME Programmatic Funding Scheme Under Project A20H4b0141.

Disclosure of Interests

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

References

1. Alsentzer, E., Murphy, J.R., Boag, W., Weng, W.H., Jin, D., Naumann, T., McDermott, M.: Publicly available clinical bert embeddings. arXiv preprint arXiv:1904.03323 (2019)
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. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swin-unet: Unet-like pure transformer for medical image segmentation. arXiv preprint arXiv:2105.05537 (2021)
4. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
5. Croitoru, F.A., Hondru, V., Ionescu, R.T., Shah, M.: Diffusion models in vision: A survey. arXiv preprint arXiv:2209.04747 (2022)
6. Degerli, A., Kiranyaz, S., Chowdhury, M.E., Gabbouj, M.: Osegnet: Operational segmentation network for covid-19 detection using chest x-ray images. arXiv preprint arXiv:2202.10185 (2022)
7. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in Neural Information Processing Systems **34**, 8780–8794 (2021)
8. Feng, C.M., Yan, Y., Fu, H., Chen, L., Xu, Y.: Task transformer network for joint mri reconstruction and super-resolution. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VI 24. pp. 307–317. Springer (2021)

9. Feng, C.M., Yu, K., Liu, Y., Khan, S., Zuo, W.: Diverse data augmentation with diffusions for effective test-time prompt tuning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2704–2714 (2023)
10. Feng, X., Yang, J., Laine, A.F., Angelini, E.D.: Discriminative localization in cnns for weakly-supervised segmentation of pulmonary nodules. In: International conference on medical image computing and computer-assisted intervention. pp. 568–576. Springer (2017)
11. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems **33**, 6840–6851 (2020)
12. Huang, S.C., Shen, L., Lungren, M.P., Yeung, S.: Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3942–3951 (2021)
13. Johnson, A.E., Pollard, T.J., Shen, L., Lehman, L.w.H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., Mark, R.G.: Mimic-iii, a freely accessible critical care database. Scientific data **3**(1), 1–9 (2016)
14. Kazerouni, A., Aghdam, E.K., Heidari, M., Azad, R., Fayyaz, M., Hacihaliloglu, I., Merhof, D.: Diffusion models for medical image analysis: A comprehensive survey. arXiv preprint arXiv:2211.07804 (2022)
15. Kim, B., Ye, J.C.: Diffusion deformable model for 4d temporal medical image generation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 539–548. Springer (2022)
16. Kumar, N., Verma, R., Anand, D., Zhou, Y., Onder, O.F., Tsougenis, E., Chen, H., Heng, P.A., Li, J., Hu, Z., et al.: A multi-organ nucleus segmentation challenge. IEEE transactions on medical imaging **39**(5), 1380–1391 (2019)
17. Li, Y., Luo, L., Lin, H., Chen, H., Heng, P.A.: Dual-consistency semi-supervised learning with uncertainty quantification for covid-19 lesion segmentation from ct images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 199–209. Springer (2021)
18. Li, Z., Li, Y., Li, Q., Zhang, Y., Wang, P., Guo, D., Lu, L., Jin, D., Hong, Q.: Lvit: language meets vision transformer in medical image segmentation. arXiv preprint arXiv:2206.14718 (2022)
19. Moghadam, P.A., Van Dalen, S., Martin, K.C., Lennerz, J., Yip, S., Farahani, H., Bashashati, A.: A morphology focused diffusion probabilistic model for synthesis of histopathology images. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2000–2009 (2023)
20. Peng, W., Adeli, E., Zhao, Q., Pohl, K.M.: Generating realistic 3d brain mris using a conditional diffusion probabilistic model. arXiv preprint arXiv:2212.08034 (2022)
21. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
22. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. arXiv preprint arXiv:2208.12242 (2022)
23. Xu, G., Song, Z., Sun, Z., Ku, C., Yang, Z., Liu, C., Wang, S., Ma, J., Xu, W.: Camel: A weakly supervised learning framework for histopathology image segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10682–10691 (2019)
24. Yu, L., Wang, S., Li, X., Fu, C.W., Heng, P.A.: Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation. In: International Confer-

ence on Medical Image Computing and Computer-Assisted Intervention. pp. 605–613. Springer (2019)