LEAF: Latent Diffusion with Efficient Encoder Distillation for Aligned Features in Medical Image Segmentation

Qilin Huang[1], Tianyu Lin[2], Zhiguang Chen[1], and Fudan Zheng[1]✉

[1] School of Computer Science and Engineering, Sun Yat-sen University, China
zhengfd5@mail.sysu.edu.cn
[2] Department of Biomedical Engineering, Johns Hopkins University, United States

Abstract. Leveraging the powerful capabilities of diffusion models has yielded quite effective results in medical image segmentation tasks. However, existing methods typically transfer the original training process directly without specific adjustments for segmentation tasks. Furthermore, the commonly used pre-trained diffusion models still have deficiencies in feature extraction. Based on these considerations, we propose LEAF, a medical image segmentation model grounded in latent diffusion models. During the fine-tuning process, we replace the original noise prediction pattern with a direct prediction of the segmentation map, thereby reducing the variance of segmentation results. We also employ a feature distillation method to align the hidden states of the convolutional layers with the features from a transformer-based vision encoder. Experimental results demonstrate that our method enhances the performance of the original diffusion model across multiple segmentation datasets for different disease types. Notably, our approach does not alter the model architecture, nor does it increase the number of parameters or computation during the inference phase, making it highly efficient. Project page: https://anonymous.4open.science/r/LEAF-F669

Keywords: Latent Diffusion · Feature Alignment · Efficient.

1 Introduction

The diffusion model[20] has achieved successful results in multiple image generation tasks, demonstrating itself as a scalable approach to generate high-dimensional visual data. Due to its powerful capabilities, recent studies have begun exploring its potential for application in other vision tasks. For example, DMP[11] adapts a text-to-image diffusion model to obtain faithful estimations on several dense prediction tasks. Marigold[10] directly fine-tunes Stable Diffusion[17] for image-conditioned depth generation, achieving state-of-the-art (SOTA) performance on multiple depth estimation datasets while enabling zero-shot transfer to unseen data.

Given the powerful capabilities of diffusion models, numerous studies have explored their application to medical image segmentation[24,25,26], demonstrating their remarkable potential and sparking growing research interest in the community. However, these methods typically directly adopt the original diffusion model training process and often incorporate overly complex modules to enhance feature representation. While enhancing performance, these designs bring about computational inefficiencies and obscure the fundamental differences between segmentation and generation tasks. In contrast, SDSeg[13] employs a latent diffusion model and improves inference speed and accuracy by utilizing a single-step reverse process. Nevertheless, this approach still fails to address the inherent divergence between segmentation objectives (e.g., pixel-wise classification) and generative modeling principles (e.g., noise prediction). Several prior works have investigated alternative parameterization methods[19,1] to generate detailed and realistic natural images. However, these approaches either rely on multi-step progressive generation or yield comparable evaluation metrics, thereby offering limited insights for this task.

Meanwhile, the widely used pre-trained diffusion models are based on a convolutional U-Net architecture. Many studies have pointed out that Transformer architectures[22] can effectively enhance feature extraction, although they also increase both the computational cost and the number of parameters. In addition, some research[12] indicates that while estimating pixel-level geometric attributes from a single image requires a comprehensive understanding of the scene, merely predicting results in the input space is insufficient for learning a robust representation. Consequently, achieving a good trade-off between speed and performance remains a significant challenge in current work. Distillation methods offer a promising solution. Notably, the recent REPA approach[27] accelerates model convergence by aligning the features of two Transformers[22], leading to improved generation performance.

Motivated by these concerns, we propose **LEAF** (**L**atent Diffusion with **E**fficient Encoder Distillation for **A**ligned **F**eatures). We analyze the diffusion formulation and discover that using noise prediction in segmentation tasks might not be optimal, as it tends to amplify prediction errors. Therefore, we replace this approach with one that directly predicts the sample. Moreover, we implement a novel and simple distillation method to enhance the feature representation of convolutional networks, allowing the model to align its features with those obtained from powerful Transformer-based models. Such alignment enhances segmentation performance without incurring any additional computational overhead or parameter increase during inference.

In conclusion, our main contributions are as follows:

- We replace the high-variance ϵ-prediction, originally used in diffusion models for generation tasks, with the low-variance x_0-prediction that is better suited for segmentation tasks, and provide the corresponding mathematical explanation.
- We design an efficient feature alignment method that enriches the representation of U-Net by distilling a powerful visual encoder, thereby improving segmentation performance on multiple medical imaging datasets across various diseases.

- Our method allows the alignment module to be removed during inference, incurring no additional computational or memory overhead. Moreover, this plug-and-play approach does not alter the internal structure of the model and can be easily transferred to other diffusion-based models.

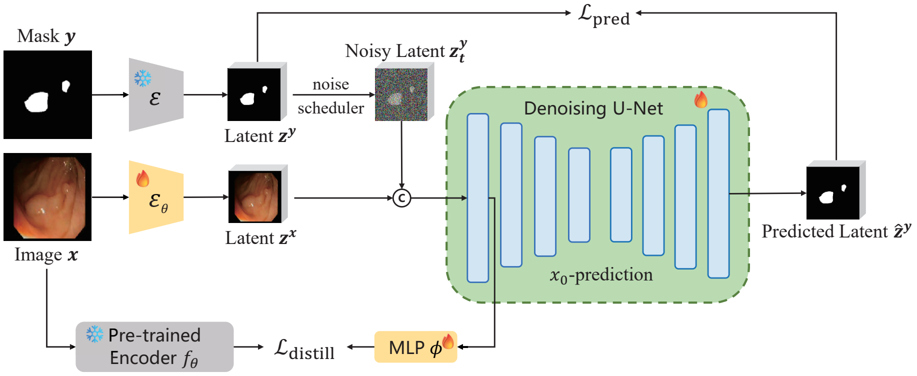
Mask $\boldsymbol{y}$
$\mathcal{E}$
Latent $\boldsymbol{z}^y$
noise
scheduler
Noisy Latent $\boldsymbol{z}_t^y$
$\mathcal{L}_{\text{pred}}$
Denoising U-Net
x_0-prediction
Predicted Latent $\hat{\boldsymbol{z}}^y$
Image $\boldsymbol{x}$
$\mathcal{E}_\theta$
Latent $\boldsymbol{z}^x$
c
Pre-trained Encoder f_θ
$\mathcal{L}_{\text{distill}}$
MLP ϕ

Fig. 1. An illustration of training pipeline of LEAF, c means concatenate, the decoder $\mathcal{D}$ is omitted and the noise scheduler is variance-preserving, e.g. $\alpha_t^2 + \sigma_t^2 {=} 1$.

2 Methods

The framework of our method is shown in Figure 1. For the conditioning approach, we follow the process of SDSeg[13]. Given a Ground-Truth segmentation map y, we use a frozen encoder $\mathcal{E}$ to map it into the latent variable z^y, and add noise via Equation (1) to obtain the noisy variable z_t^y, where ϵ is random Gaussian noise, and α_t and σ_t are a set of hyperparameters pre-defined by the noise scheduler; typically, α_t decreases while σ_t increases. For the image x, we use a learnable encoder $\mathcal{E}_\theta$, initialized with the weights of $\mathcal{E}$, to map it into z^x. Then, we concatenate $\text{concat}(z^x; z_t^y)$ as the input to the U-Net and obtain the output $\hat{z}^y$.

$$z_t = \alpha_t z + \sigma_t \epsilon \tag{1}$$

2.1 Parameterization Types

Current diffusion models are used primarily for generation tasks, and prediction objectives generally utilize the following two approaches: (1) ϵ-prediction[9], in which the model learns to predict the noise ϵ; (2) $\mathbf{v}$-prediction[19], in which the model learns to predict the velocity defined by Equation (2).

$$\mathbf{v} := \alpha_t \epsilon - \sigma_t z \tag{2}$$

These parameterization types can be used to estimate the original image via the formula in Equation (3):

$$\hat{z} = \begin{cases} (z_t - \sigma_t \hat{\epsilon})/\alpha_t & \text{, If } \epsilon\text{-prediction} \\ \alpha_t z_t - \sigma_t \hat{\mathbf{v}} & \text{, If } \mathbf{v}\text{-prediction} \\ \hat{z} & \text{, If } x_0\text{-prediction} \end{cases} \tag{3}$$

Since we freeze the decoder $\mathcal{D}$ during training, the error in $\hat{x}$ originates primarily from $\hat{z}$. Note that both parameterization methods involve a coefficient based on variance when reconstructing z_t; as $t \to T$, σ_t increases while α_t decreases. In diffusion models that employ a single-step reverse process at $t = T$, this further amplifies the error in the estimation.

Therefore, we propose that the x_0-prediction[9] approach is more suitable for image segmentation. Using x_0-prediction, the diffusion model directly outputs $\hat{z}$ without introducing additional scaling coefficients, thereby avoiding unnecessary errors. Compared to the other two prediction methods, this approach yields more stable and accurate results. In summary, for a diffusion model pre-trained with either ϵ-prediction or $\mathbf{v}$-prediction, we fine-tune it directly using x_0-prediction so that it directly predicts the segmentation map. The corresponding loss is shown in the Equation (4), using the $L1$ loss as implemented in SDSeg.

$$\mathcal{L}_{\text{pred}} = \mathcal{L}_{L1}(\hat{z}^y, z^y) \tag{4}$$

2.2 Features Alignment

In recent studies, enhancing the ability of diffusion models to extract features is typically achieved by modifying the model architecture. As demonstrated in TransUNet[4] and Diff-Trans[5], Transformer architectures can effectively improve the encoder's feature extraction capability, but they also significantly increase the number of model parameters and computational cost. To encourage a U-Net based latent diffusion model to learn rich representations, we introduce a regularization strategy to augment the capacity of convolution, enabling it to capture the representations learned by Transformer architectures.

Inspired by REPA[27], we utilize a pre-trained self-supervised powerful visual encoder f_θ, such as DINOv2[14] or CLIP[16], as the base model for providing robust visual representations. It takes a clean image x as input and produces hidden states $h = f_\theta(x) \in \mathbb{R}^{L \times D}$, where L is the number of patches and D is the embedding dimension. In the encoder block of the denoising U-Net, we obtain a feature map $m \in \mathbb{R}^{C \times H \times W}$ and reshape it into $\mathbb{R}^{HW \times C}$, with the condition that $HW = L$. Then, we use a multilayer perceptron (MLP) ϕ to project m, yielding $\phi(m) \in \mathbb{R}^{L \times D}$, and compute a distillation loss based on cosine similarity:

$$\mathcal{L}_{\text{distill}} = \mathcal{L}(h, \phi(m)) = -\frac{1}{N} \sum_{i=1}^{N} \left(\frac{h_i \cdot m_i}{\|h_i\| \cdot \|m_i\|}\right) \tag{5}$$

We add this loss to the prediction loss above, and the total objective loss is shown in Equation(5), where λ is a positive constant controlling the strength of the distillation alignment.

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{pred}} + \lambda \mathcal{L}_{\text{distill}} \tag{6}$$

2.3 Inference

During inference, we initialize the segmentation map with standard Gaussian noise z_T^y, and encode the input image into z^x using $\mathcal{E}_\theta$. The concatenated features $(z_T^y; z^x)$ are then fed into the U-Net. Notably, we remove the pre-trained visual encoder and MLP during this phase, ensuring no additional parameters are introduced compared to the original model. Following SDSeg[13], we apply a single-step reverse process to obtain $\hat{z}^y$, which is then decoded to the pixel space via $\mathcal{D}$ to produce the final segmentation map.

3 Experimental Results

3.1 Experimental setup

Datasets and Evaludation Metrics To comprehensively evaluate the proposed method, we conduct experiments on four public medical image segmentation tasks: (1) Optic-cup segmentation from retinal fundus images (REFUGE2 (REF)[15]), (2) Polyp segmentation from colonoscopy images (CVC-ClinicDB (CVC)[2]), (3) COVID-19 lesion segmentation (QaTa-Covid19 (Qata)[7]), and (4) Skin lesion segmentation from dermoscopy images (ISIC 2018[6,21]). We use mean Dice and mean IoU as primary evaluation metrics. For REFUGE2, we used the data partition defined in SDSeg. For ISIC 2018, we adopted a training-testing ratio of 7 : 3, while for CVC, we utilized an 80:10:10 data partition. QaTa, on the other hand, used the default training and testing sets.

Implementation Details We implemented LEAF using the PyTorch platform and trained/evaluated on a single NVIDIA A800 GPU. All images were resized to a resolution of 256 × 256. The pre-trained unconditional latent diffusion model was based on LDM-KL-8[17]. To optimize the model, we utilized the standard AdamW optimizer with a batch size of 4. The learning rate was set to 4×10^{-5} with a warm-up constant learning rate scheduler. To handle the concanated input, we duplicated the U-Net input layer from 4 channels to 8 channels and initialized its weights by halving the original weights as mentioned in Marigold[10]. The pre-trained vision encoder was DINOv2[14].

3.2 Performance Comparison

We conducted extensive experiments in various evaluated datasets to assess the effectiveness of LEAF, as shown in Table 1. LEAF represented a generic approach for latent diffusion models without domain-specific modules tailored to particular medical imaging modalities. Consequently, our comparisons focused on models with strong generalization capabilities. For a fair comparison with SDSeg, we re-trained it under our framework using the same configurations. MedSegDiff was re-evaluated on CVC. The results of QaTa and ISIC2018 were directly cited from MMDSN[23] and BGDiffSeg[8], and REF from SDSeg[13], CVC from Diff-Trans[5].

Table 1. Performance comparison on our proposed model and existing SOTA medical segmentation models.

	Model	REF		CVC		QaTa		ISIC2018	
		Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU
CNN/Transformer-based	U-Net[18]	80.1	-	85.6	80.5	79.0	69.5	87.6	77.9
	TransUNet[4]	85.6	-	92.0	87.8	78.6	69.1	88.7	79.7
	Swin-UNet[3]	84.3	-	91.4	87.4	78.1	68.3	-	-
Diffusion-based	MedSegDiff[24]	86.3	78.2	92.4	88.9	76.5	67.2	85.5	74.7
	SDSeg[13]	88.7	80.9	93.6	89.3	77.6	68.0	88.1	79.7
Proposed	LEAF	**89.5**	**81.5**	**95.2**	**90.9**	**80.2**	**71.0**	**90.5**	**84.1**

As shown in Table 1, LEAF outperforms all baseline models on datasets involving various types of medical image, validating its effectiveness and generalizability. While sharing the same U-Net architecture as SDSeg, our method replaces its parameterization type and aligns convolutional layers with features extracted from a Transformer-based encoder, achieving significant performance improvements.

3.3 Ablation Study

Ablation for fine-tuning pipeline We establish the baseline as the model using the original ϵ-prediction without feature alignment (first row in Table 2). From the table, we observe that merely changing the prediction method from ϵ-prediction to $\mathbf{v}$-prediction yields significant performance gains. Furthermore, switching to x_0-preditcion without scaling factors further improves model performance. Finally, feature alignment achieves the highest performance compared to other configurations. Although these features come from different model architectures and DINOv2 is not fine-tuned on medical images, it can still improve the segmentation performance. We emphasize that these improvements are consistent across all evaluated datasets, with nontrivial magnitude.

Ablation for Feature Alignment We investigate the hyperparameter λ that controls alignment strength and the model size of the vision encoder, with results shown in Table 3. Firstly, the optimal value of λ generally varies across different datasets, which may be related to the distribution and inherent characteristics

Table 2. Ablation study for parameterization and features alignment. The rows with gray color highlight the features of the model distilled from the vision encoder during training for clearer comparison.

Parameterization	Feature	REF		CVC		QaTa		ISIC2018	
Types	Alignment	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU
ϵ-prediction	✗	88.47	79.59	90.15	83.68	74.27	63.80	87.67	80.13
ϵ-prediction	✓	87.61	78.43	91.63	87.10	74.56	63.97	87.34	79.48
$\mathbf{v}$-prediction	✗	89.21	80.92	93.75	89.32	79.10	69.74	90.35	83.91
$\mathbf{v}$-prediction	✓	89.30	80.88	94.89	90.44	79.32	69.98	90.52	84.11
x_0-prediction	✗	89.21	79.08	94.49	89.94	79.08	69.85	90.39	83.94
x_0-prediction	✓	89.53	81.45	95.17	90.94	80.15	71.04	90.54	84.18

of the data. Secondly, the impact of different λ values on model performance is not significant, for example, the maximum absolute difference of dice score on ISIC2018 is 0.3, while on QaTa, it is over 1.0. We believe this is due to the fact that the images in QaTa contain more structural information, making them more sensitive to the distillation strength. Overall, the model performs better with $\lambda > 0$ than it does with $\lambda = 0$, further validating the effectiveness of feature alignment.

Table 3. The effect of hyperparameter λ for features alignment.

λ	REF		CVC		QaTa		ISIC2018	
	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU
0	89.21	79.08	94.49	89.94	79.08	69.85	90.39	83.94
0.15	89.39	81.19	95.07	90.77	79.62	70.37	90.24	83.83
0.25	89.41	81.24	94.97	90.57	79.74	70.65	**90.54**	**84.06**
0.50	89.33	81.12	94.21	89.35	80.05	70.87	90.43	84.06
0.75	**89.53**	**81.45**	95.01	90.67	79.98	70.93	90.51	84.05
1.0	89.44	81.27	**95.17**	**90.94**	79.87	70.77	90.34	83.90
1.25	89.43	81.27	95.07	90.76	**80.15**	**71.04**	90.30	83.93

3.4 Quality Results

Stability Diffusion models are non-deterministic models; thus, many previous models have attempted to reduce segmentation uncertainty by running the model multiple times and ensembling the results in some way as the final outcome, which increases the inference speed of the model. The segmentation results for LEAF are obtained from a single run, so it is necessary to demonstrate their stability, with the results presented in Table 4.

The experimental results demonstrated that the ϵ-prediction method indeed has a larger variance compared to the x_0-prediction method. Moreover, the standard deviation of the models trained using the x_0-prediction method is mostly

Table 4. Stability experiments. We selected 10 different random seeds to inference 10 times and calculated the standard deviation of the Dice score (%).

Parameterization Types	REF	CVC	QaTa	ISIC2018
ϵ-prediction	0.09	0.32	0.14	0.29
x_0-prediction	0.05	0.11	0.08	0.06

on the order of 10^{-2}, indicating a very small difference. This difference is significantly smaller than the improvement brought about by our proposed method, further proving the effectiveness and stability of our approach.

Visualization Additionally, we visualize the segmentation results on different medical segmentation tasks, as shown in Figure 2.

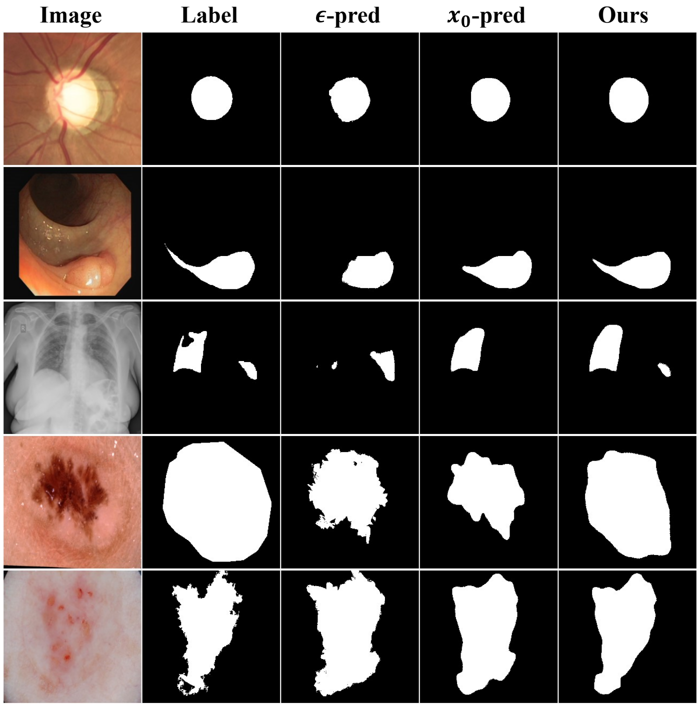

Fig. 2. Visualization of segmentation results.

4 Conclusion

In this paper, we propose LEAF, an efficient and generalized framework for fine-tuning a latent diffusion model for medical image segmentation. We investigate the effect of parameterization type and propose to use x_0-prediction parameterization for segmentation task. We also introduce a simple featrue alignment method via distilling vision encoder, providing a better representation for the CNN-based U-Net. LEAF brings about zero cost for inference and can be easily adapted to other LDM-based segmentation models.

Acknowledgments. This study was funded by the Guangdong S&T Program (Grant No. 2024B0101040005), the National Natural Science Foundation of China (Grant No. 62461146204), the National Key Research and Development Program of China (Grant No. 2021YFB0300103), and the Pazhou Lab program (Grant No. PZL2023KF0001.)

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

References

1. Benny, Y., Wolf, L.: Dynamic dual-output diffusion models. In: CVPR. pp. 11472–11481. IEEE (2022)
2. Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., Gil, D., de Miguel, C.R., Vilariño, F.: WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. Comput. Medical Imaging Graph. **43**, 99–111 (2015)
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. In: ECCV Workshops (3). Lecture Notes in Computer Science, vol. 13803, pp. 205–218. Springer (2022)
4. Chen, J., Mei, J., Li, X., Lu, Y., Yu, Q., Wei, Q., Luo, X., Xie, Y., Adeli, E., Wang, Y., et al.: Transunet: Rethinking the u-net architecture design for medical image segmentation through the lens of transformers. Medical Image Analysis p. 103280 (2024)
5. Chowdary, G.J., Yin, Z.: Diffusion transformer u-net for medical image segmentation. In: MICCAI (4). Lecture Notes in Computer Science, vol. 14223, pp. 622–631. Springer (2023)
6. Codella, N.C.F., Gutman, D.A., Celebi, M.E., Helba, B., Marchetti, M.A., Dusza, S.W., Kalloo, A., Liopyris, K., Mishra, N.K., Kittler, H., Halpern, A.: Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (ISIC). In: ISBI. pp. 168–172. IEEE (2018)
7. Degerli, A., Kiranyaz, S., Chowdhury, M.E.H., Gabbouj, M.: Osegnet: Operational segmentation network for covid-19 detection using chest x-ray images. In: ICIP. pp. 2306–2310. IEEE (2022)
8. Guo, Y., Cai, Q.: BGDiffSeg: a Fast Diffusion Model for Skin Lesion Segmentation via Boundary Enhancement and Global Recognition Guidance . In: proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. vol. LNCS 15009. Springer Nature Switzerland (October 2024)
9. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems. vol. 33, pp. 6840–6851. Curran Associates, Inc. (2020)

10. Ke, B., Obukhov, A., Huang, S., Metzger, N., Daudt, R.C., Schindler, K.: Repurposing diffusion-based image generators for monocular depth estimation. In: CVPR. pp. 9492–9502. IEEE (2024)
11. Lee, H., Tseng, H., Lee, H., Yang, M.: Exploiting diffusion prior for generalizable dense prediction. In: CVPR. pp. 7861–7871. IEEE (2024)
12. Li, A.C., Prabhudesai, M., Duggal, S., Brown, E., Pathak, D.: Your diffusion model is secretly a zero-shot classifier. In: ICCV. pp. 2206–2217. IEEE (2023)
13. Lin, T., Chen, Z., Yan, Z., Yu, W., Zheng, F.: Stable diffusion segmentation for biomedical images with single-step reverse process. In: MICCAI (8). Lecture Notes in Computer Science, vol. 15008, pp. 656–666. Springer (2024)
14. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P., Li, S., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jégou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision. Trans. Mach. Learn. Res. **2024** (2024)
15. Orlando, J.I., Fu, H., Breda, J.B., van Keer, K., Bathula, D.R., Diaz-Pinto, A., Fang, R., Heng, P., Kim, J., Lee, J., Lee, J., Li, X., Liu, P., Lu, S., Murugesan, B., Naranjo, V., Phaye, S.S.R., Shankaranarayana, S.M., Bogunovic, H.: REFUGE challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. Medical Image Anal. **59** (2020)
16. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: ICML. Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763. PMLR (2021)
17. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR. pp. 10674–10685. IEEE (2022)
18. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI (3). Lecture Notes in Computer Science, vol. 9351, pp. 234–241. Springer (2015)
19. Salimans, T., Ho, J.: Progressive distillation for fast sampling of diffusion models. In: ICLR. OpenReview.net (2022)
20. Sohl-Dickstein, J., Weiss, E.A., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: ICML. JMLR Workshop and Conference Proceedings, vol. 37, pp. 2256–2265. JMLR.org (2015)
21. Tschandl, P., Rosendahl, C., Kittler, H.: Descriptor : The ham 10000 dataset , a large collection of multi-source dermatoscopic images of common pigmented skin lesions (2018), https://api.semanticscholar.org/CorpusID:263789934
22. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: NIPS. pp. 5998–6008 (2017)
23. Wang, H., Zhang, Z., Zhang, Y., Zhang, J., Ou, Y., Sun, X.: Multi-modal diffusion network with controllable variability for medical image segmentation. In: 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp. 3817–3822 (2024). https://doi.org/10.1109/BIBM62325.2024.10822810
24. Wu, J., Fang, H., Zhang, Y., Yang, Y., Xu, Y.: Medsegdiff: Medical image segmentation with diffusion probabilistic model. In: International Conference on Medical Imaging with Deep Learning (2022)
25. Wu, J., Ji, W., Fu, H., Xu, M., Jin, Y., Xu, Y.: Medsegdiff-v2: Diffusion-based medical image segmentation with transformer. In: AAAI. pp. 6030–6038 (2024)

26. Xing, Z., Wan, L., Fu, H., Yang, G., Zhu, L.: Diff-unet: A diffusion embedded network for volumetric segmentation. CoRR **10326** (2023)
27. Yu, S., Kwak, S., Jang, H., Jeong, J., Huang, J., Shin, J., Xie, S.: Representation alignment for generation: Training diffusion transformers is easier than you think. In: International Conference on Learning Representations (2025)