Improving Medical Image Generative Models with Fréchet Distance Loss

Andrew Marshall[1], Xuanang Xu[2], Xiaoran Zhang[1], Rui Wang[1], Lawrence H. Staib[1,2,3], and James S. Duncan[1,2,3]

[1] Department of Biomedical Engineering, Yale University,
[2] Department of Radiology & Biomedical Imaging, Yale University,
[3] Department of Electrical Engineering, Yale University,
andrew.marshall@yale.edu

Abstract. Diffusion generative models have demonstrated immense potential for synthetic medical image generation. However, these models often struggle to capture complex morphological characteristics of heterogeneous tumors with irregular boundaries, limiting their utility for downstream clinical tasks such as segmentation. This limitation stems from the standard denoising objective: minimizing a per-pixel error, which smooths high-variance irregular structures characteristic of tumors. To address this, we propose finetuning these generative models with Fréchet Distance loss (FD-loss). FD-loss aligns the first and second order feature statistics of real and generated images in a pretrained encoder space, encouraging the generator to capture complex structural variations characteristic of heterogeneous tumors. We integrate FD-loss across diverse architectural settings, using both natural- and medical-image encoders on multiple liver and brain cancer datasets spanning CT and MRI modalities. Downstream segmentation networks trained on our FD-regularized synthetic data consistently achieve superior performance, improving tumor DSC by up to 5% over unregularized synthetic augmentation alone. Qualitative analysis suggests these gains are associated with more faithful tumor synthesis and fewer segmentation hallucinations. Our results show FD-loss as an effective regularizer for medical image generative models to improve clinical workflows.

Keywords: Generative Models · Segmentation · Fréchet Distance.

1 Introduction

Generative models have emerged as a promising paradigm for synthesizing medical image data. Recent methods, ranging from unconditional medical image generation to segmentation-guided diffusion [9,25] and joint image-mask pair synthesis [12], have shown potential for producing high-fidelity synthetic data to augment low-data regimes. However, these models often struggle to capture structures with large variations in size, shape, and other morphological characteristics [9]. Such failures can yield unrealistic synthetic images that, when used

for downstream task training (e.g., data augmentation for tumor segmentation), degrade model performance on real clinical samples.

During generative model training, image quality is commonly assessed using distributional metrics. Fréchet Inception Distance (FID) measures the similarity between image distributions using features extracted from an InceptionV3 encoder trained on natural images [6,21]. However, FID does not always correlate with medical image generation quality [9]. In the medical imaging domain, RadInceptionV3 replaces the natural-image encoder with an InceptionV3 model pretrained on RadImageNet, a dataset of 1.35 million annotated CT, MRI, and ultrasound exams [14]. This substitution enables medical images to be analyzed in a more domain-appropriate feature space with RadFID [22]. Likewise, Fréchet Radiomic Distance (FRD) [10] compares image distributions using radiomic features. Notably, segmentation accuracy on translated medical images has been shown to correlate with both RadFID [22] and FRD [10]. Given this relationship, it is natural to ask whether Fréchet distance can serve not only as an evaluation metric but also as a loss function for improving generative models designed to support segmentation.

FID has recently been introduced as a loss function for training generative models on natural images [13,24]. Previously, using FID as a loss was impractical because reliable Fréchet Distance (FD) estimation required large sample sizes, often on the order of 50,000 images. Generating this many samples during training is empirically infeasible for small medical imaging datasets. Yang et al. [24] addressed this limitation with a decoupled optimization framework in which distribution statistics are precomputed from ground-truth images. They used a small queue of generated images ($n = 5,000$) to provide a useful FD signal during optimization. Whereas prior work has focused on optimizing FID for general image generators, its utility for medical image generation remains unexplored. In this study, we investigate whether FD-loss can improve segmentation-guided generative models by encouraging more realistic synthetic medical images. Specifically, we finetune generative models with FD-loss computed across multiple embedding spaces defined by different pretrained image encoders, including InceptionV3, RadInceptionV3, BiomedCLIP, and MedDINOv3. We then evaluate the resulting synthetic images through downstream medical image segmentation experiments. Our contributions are threefold: (1) we adapt queue-based FD-loss for finetuning segmentation-guided medical diffusion models; (2) we evaluate multiple natural, radiological, and biomedical feature spaces for optimization; (3) we demonstrate on liver CT and brain MRI datasets that FD-regularized synthetic augmentation outperforms traditional strategies.

2 Preliminaries

2.1 Segmentation-Guided Generative Models

Segmentation-guided diffusion models generate medical images by conditioning the denoising process on an anatomical segmentation mask [9]. Let $x_0 \in \mathbb{R}^{c \times h \times w}$ denote a medical image and let $m \in \{0, \ldots, C-1\}^{h \times w}$ denote a multi-class

anatomical mask. Rather than modeling the unconditional image distribution $p(x_0)$, the segmentation-guided model learns the conditional distribution $p(x_0 \mid m)$ so that generated images follow the spatial anatomy specified by the input mask. The forward noising process is unchanged from a standard diffusion model, $x_t = \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon$, with $\epsilon \sim \mathcal{N}(0, I)$ [7]. Rather than predicting the added noise ϵ, we adopt the velocity v parameterization [18], in which the network regresses the velocity target $v_t = \sqrt{\bar{\alpha}_t}\, \epsilon - \sqrt{1-\bar{\alpha}_t}\, x_0$. In practice, the mask is concatenated channel-wise with the noisy image and passed to a UNet, yielding the velocity estimate $v_\theta(x_t, t \mid m)$. The model is trained with the mask-conditional objective $\mathcal{L}_m = \mathbb{E}_{(x_0,m),t,\epsilon}\big[\|v_t - v_\theta(\sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon,\, t \mid m)\|^2\big]$. Because the mask is supplied throughout the iterative denoising trajectory, the generator receives explicit spatial guidance for anatomical structures while retaining stochastic variation in image appearance. The clean image can be recovered from the predicted velocity via $\hat{x}_0 = \sqrt{\bar{\alpha}_t}\, x_t - \sqrt{1-\bar{\alpha}_t}\, v_\theta(x_t, t \mid m)$, and sampling is performed with an accelerated sampler such as DDIM [20] to synthesize images anatomically aligned with the conditioning mask.

2.2 Fréchet Distance

Fréchet distance compares two image sets using statistics computed from their feature representations. The real image set $\mathcal{R} = \{\mathbf{x}_i\}$ and the generated image set $\mathcal{G} = \{\hat{\mathbf{x}}_i\}$ are passed through a feature extractor $\phi(\cdot)$, and the resulting feature distributions are approximated as multivariate Gaussians with means $\boldsymbol{\mu}_r = \mathbb{E}[\phi(\mathbf{x})]$, $\boldsymbol{\mu}_g = \mathbb{E}[\phi(\hat{\mathbf{x}})]$ and covariances $\boldsymbol{\Sigma}_r = \mathrm{Cov}[\phi(\mathbf{x})]$, $\boldsymbol{\Sigma}_g = \mathrm{Cov}[\phi(\hat{\mathbf{x}})]$. The distance between these two Gaussians combines the squared difference between their means with a covariance-matching term:

$$\mathrm{FD}_\phi(\mathcal{R}, \mathcal{G}) = \|\boldsymbol{\mu}_r - \boldsymbol{\mu}_g\|_2^2 + \mathrm{Tr}\Big(\boldsymbol{\Sigma}_r + \boldsymbol{\Sigma}_g - 2\,(\boldsymbol{\Sigma}_r \boldsymbol{\Sigma}_g)^{\frac{1}{2}}\Big). \tag{1}$$

When ϕ is chosen as InceptionV3 [21], this formulation reduces to the familiar Fréchet Inception Distance (FID) [6]. When FD is used as an evaluation metric, the real-image statistics $(\boldsymbol{\mu}_r, \boldsymbol{\Sigma}_r)$ are computed once over the training set, whereas the generated-image statistics $(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$ are estimated from a large pool of samples, often numbering in the tens of thousands.

3 Methods

3.1 Finetuning Generative Models with Fréchet Distance Loss

Following the FD-loss framework [24], we finetune a pretrained segmentation-guided generator by minimizing the Fréchet distance between real and generated image distributions in a frozen representation space. For a chosen feature extractor $\phi(\cdot)$, the real-image statistics $(\boldsymbol{\mu}_r, \boldsymbol{\Sigma}_r)$ are precomputed once from the training set and held fixed during optimization. Before finetuning begins, the pretrained generator synthesizes $N = 5{,}000$ images from real segmentation masks

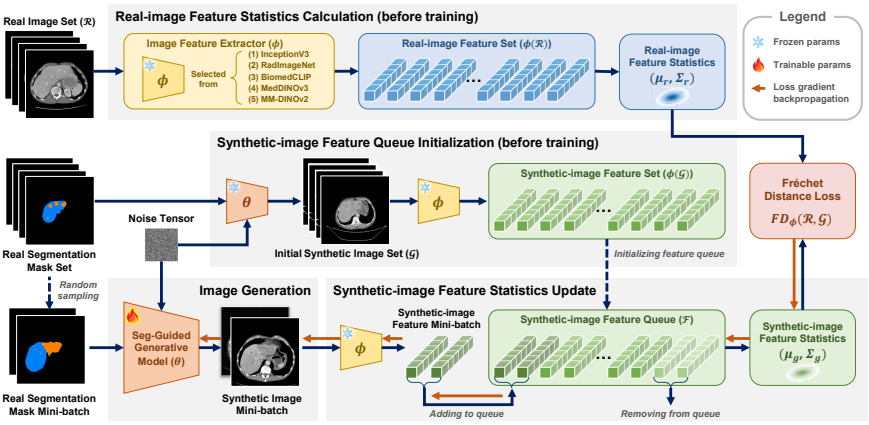

Fig. 1. Our proposed segmentation-guided generative model finetuning with FD-loss.

using 25-step DDIM sampling. These images are passed through the frozen extractor to obtain N synthetic-image features, which initialize a feature queue $\mathcal{F}$ of fixed size N used to estimate the generated-image distribution.

During finetuning, each iteration samples a mini-batch of $M = 16$ segmentation masks and synthesizes M images through a differentiable 25-step DDIM trajectory, retaining gradients through the full sampling chain. To accelerate queue turnover without incurring the memory cost of backpropagating through additional sampling trajectories, we simultaneously generate a further 32 images per iteration without gradient tracking; these images are also passed through $\phi(\cdot)$ and added to the queue, but contribute only to the generated-image statistics, not to the loss gradient. All 48 new features per iteration are added to $\mathcal{F}$, and the oldest 48 features are removed, keeping the queue size fixed at N. The generated statistics $(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$ are then computed from this updated queue, and the FD between the fixed real statistics and the queue-based generated statistics is used as the training objective (Eq. 1). Gradients are propagated only through the $M = 16$ gradient-tracked images and the generator parameters, while all other queue features remain detached. This decouples the population size needed for stable FD estimation from the batch size used for gradient computation, making FD practical as a finetuning loss for medical image generators. We finetune each generator for 20,000 steps using a learning rate of 1e−7, monitoring the FD-loss trajectory to confirm convergence, on a single NVIDIA H200 GPU.

3.2 Training Segmentation Models with Synthetic Data

To evaluate the realism and downstream utility of our synthetic data, we trained segmentation models and evaluated their performance on real clinical scans. We trained a standard 2D U-Net [16] for 100 epochs using a batch size of 8, an initial learning rate of 10^{-3}, and a cosine annealing schedule with the MONAI [5] framework. For each run, we retained the model weights that achieved the lowest

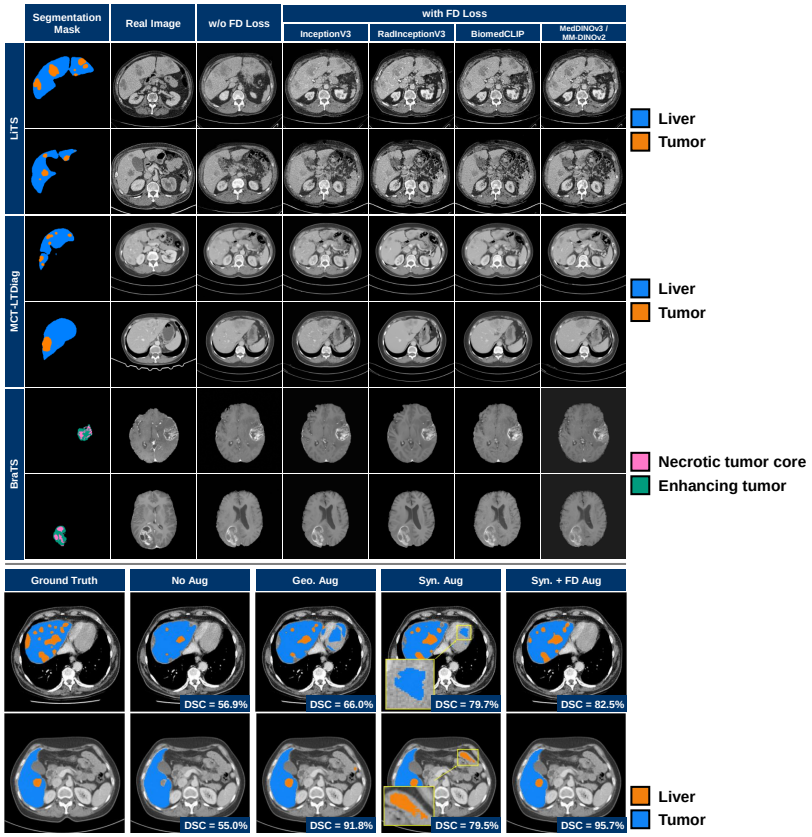

Fig. 2. Qualitative evaluation of FD-loss finetuning for image generation and downstream segmentation. Top: segmentation-conditioned synthetic images generated with different encoder spaces. Bottom: downstream liver CT segmentation results.

validation loss. To mitigate the run-to-run variance inherent to single-seed U-Net training, we trained an ensemble of five models for each configuration. We used a fixed set of random seeds across all experiments so that every configuration was evaluated under identical initializations, ensuring that performance differences are attributable to FD-loss regularization. During inference, class probabilities predicted by the five models were averaged, and a pixel-wise argmax was applied to obtain the final segmentation mask. All downstream quantitative results are reported using this soft-voting ensemble approach.

Table 1. Synthetic image quality before and after FD-loss finetuning. Best **bold**.

Models	FID ↓	KID ↓	IS ↑	CMMD ↓	FRD ↓
LiTS	99.20	0.1020	1.78	0.5915	63.56
LiTS+FD	**61.37**	**0.0526**	**1.93**	**0.1861**	**24.05**
MCT	20.40	0.0090	1.91	0.0740	69.19
MCT+FD	**18.15**	**0.0061**	**1.95**	**0.0329**	**4.37**
BraTS	38.32	**0.0294**	1.61	0.3899	58.45
BraTS+FD	**36.67**	0.0298	**1.77**	**0.2683**	**5.98**

Table 2. Downstream segmentation with different data augmentations. Best results **bold**, second best underlined.

Data	Aug.	DSC [%] ↑		ASSD [mm] ↓	
		Liv.	Tum.	Liv.	Tum.
LiTS	No aug.	**88.5**	14.5	**4.5**	23.9
	Geo.	85.9	44.9*	13.1	14.2*
	Syn.	86.6	46.9*	10.0	13.9*
	Syn.+FD	87.8	**52.3**‡	10.6	**13.7**‡
MCT	No aug.	88.2	45.9	**5.7**	10.4
	Geo.	**88.3**	60.8*	6.0	**8.9***
	Syn.	87.9	60.9*	6.7	9.8*
	Syn.+FD	88.1	**64.5**‡	7.4	9.9*
		NCR	**ET**	**NCR**	**ET**
BraTS	No aug.	36.3	61.4	9.4	7.7
	Geo.	**47.6***	**68.2***	8.8*	5.7*
	Syn.	46.4*	65.0*	8.1*	6.0*
	Syn.+FD	47.4*	66.9*	**7.3**‡	**4.9**‡

* denotes significance vs. no augmentation;
‡ denotes significance vs. both no augmentation and geometric augmentation.

Table 3. Feature-extractor ablation.

Data	Encoder	DSC [%] ↑		ASSD [mm] ↓	
		Liv.	Tum.	Liv.	Tum.
LiTS	InceptionV3	83.5	44.5	11.6	20.4
	RadInceptionV3	86.1	47.2	13.1	19.3
	BiomedCLIP	86.7	47.7*	11.7	16.6
	MedDINOv3	**87.8***	**52.3***	**10.6**	**13.7***
MCT	InceptionV3	87.5	55.8	7.5	10.5
	RadInceptionV3	82.8	52.4	10.3	22.2
	BiomedCLIP	87.6	62.9*	7.4	10.5
	MedDINOv3	**88.1**	**64.5***	**7.4**	**9.9***
		NCR	**ET**	**NCR**	**ET**
BraTS	InceptionV3	42.9	65.3	8.0	5.9
	RadInceptionV3	38.8	63.3	8.3	6.3
	BiomedCLIP	**47.4***	**66.9***	**7.3***	**4.9***
	MM-DINOv2	46.0	65.5*	9.9	6.8

* denotes statistical significance relative to the unregularized segmentation-guided synthetic-augmentation baseline.

Table 4. Distance-function ablation.

Method	DSC [%] ↑		ASSD [mm] ↓	
	Liv.	Tum.	Liv.	Tum.
FD μ Only	81.8*	27.3*	14.4*	27.1*
FD Σ Only	86.7*	41.7*	11.9*	18.5*
Symmetric KL	86.9*	38.8*	11.6*	21.2*
Bhattacharyya	86.8*	44.2*	**10.5**	17.1*
FD	**87.8**	**52.3**	10.6	**13.7**

* denotes statistical significance relative to FD baseline.

4 Experiments

4.1 Datasets and Metrics

We evaluate our method on two contrast-enhanced liver CT datasets and one multi-parametric brain MRI dataset. **Liver Tumor Segmentation (LiTS)** [3] CT dataset contains both primary and secondary hepatic tumors. Following the experimental design of Konz et al. [9], we partition the data at the patient level into three mutually exclusive subsets: 100 patients for generative model training (∼13,250 images), 16 patients for generative testing (∼3,000 images), and an additional held-out set of 15 patients (∼2,600 images) reserved for downstream evaluation. **Multi-phase CT dataset for Liver Tumor Diagnosis (MCT-LTDiag)** [23] comprises 517 patients spanning HCC, ICC, and colorectal and breast-cancer liver metastases. We partition this dataset into 302 patients (∼12,890 images) for generative model training, 110 patients (∼2,000 images) for

generative testing, and 105 patients (∼2,000 images) for downstream evaluation. For both CT datasets, volumes are windowed to a Hounsfield unit (HU) range of $[-200, 250]$, and axial slices are saved at 256×256 pixels and normalized to $[0, 1]$. To assess generalization beyond abdominal CT, we further evaluate on the **Brain Tumor Segmentation (BraTS)** dataset [15,1,2] containing multi-parametric brain MRI with glioma sub-regions. For the BraTS dataset 260 patients (∼11,841 images) were used for generative model training, 59 (∼2,598 images) were used for generative testing, and 50 (∼2,305 images) were used for downstream task evaluation. We generate T1Gd images conditioned on enhancing tumor (ET) and necrotic core (NCR) labels. Each held-out downstream evaluation set was divided in half by patient into a validation set and a test set. The validation set was used for model selection; all reported downstream results were computed on the independent test set.

We evaluate synthetic image quality using distributional metrics that compare generated and real images: Fréchet Inception Distance (FID) [6], Kernel Inception Distance (KID) [4], Inception Score (IS) [17], CLIP Maximum Mean Discrepancy (CMMD) [8], and Fréchet Radiomic Distance (FRD) [10]. Downstream segmentation utility is assessed by training segmentation models under each synthetic-data configuration and testing them on held-out real images using the Dice similarity coefficient (DSC) and average symmetric surface distance (ASSD). Statistical significance between augmentation configurations was assessed using a Wilcoxon signed-rank test. Significance is reported at $p<0.05$.

4.2 Synthetic Image Quality Evaluation

We first evaluate whether FD-loss directly improves generative model performance by enhancing synthetic image quality. For each dataset, we compare synthetic images produced by the baseline segmentation-guided generative model trained without FD-loss with images produced by the same model finetuned with FD-loss. Both models are sampled using the same conditioning masks, and the resulting synthetic image distributions are compared with the corresponding real-image distributions using the image-quality metrics described above (MedDINOv3 for liver CT, BiomedCLIP for BraTS; see Table 3). The quantitative results in Table 1 demonstrate the effectiveness of FD-loss: after finetuning, generated samples generally move closer to the real data distribution across LiTS, MCT-LTDiag, and BraTS, as reflected by lower FID, KID, CMMD, and FRD scores and an IS closer to the real-image reference. Representative qualitative examples are shown in Fig. 2.

4.3 Downstream Segmentation with Synthetic Data Augmentation

We next evaluate whether FD-regularized generative models improve downstream segmentation when used as an advanced data augmentation strategy. This experiment provides more direct evidence for the significance of FD-loss in generative model finetuning, because the primary purpose of synthetic images

is to improve downstream model training rather than merely to improve image-level metrics. If segmentation models trained with FD-regularized synthetic data outperform models trained with conventional augmentation alone, this demonstrates both the effectiveness and practical value of the proposed approach. We therefore train segmentation models under four progressively stronger settings: (1) no data augmentation; (2) conventional augmentation only, including random scaling, rotation, flipping, deformation, and intensity rescaling; (3) augmentation with synthetic images generated by the baseline model trained without FD-loss; and (4) augmentation with synthetic images generated by the model finetuned with FD-loss. The results in Table 2 show that FD-regularized synthetic images ("**Syn.+FD**") generally improve segmentation performance relative to the unaugmented ("**No aug.**"), conventionally augmented ("**Geo.**"), and unregularized synthetic-data ("**Syn.**") baselines, demonstrating the practical utility of FD-loss for downstream segmentation training.

4.4 Ablation Study

Image Feature Extractor We further perform an encoder-selection experiment to justify the choice of image feature extractor $\phi(\cdot)$ in the FD-loss framework (Fig. 1). Because FD-loss optimizes the generator in the representation space defined by $\phi(\cdot)$, different feature extractors may emphasize different anatomical, textural, or modality-specific properties of the generated images. We therefore finetune the generative models using FD-loss computed with multiple candidate feature extractors. Specifically, we consider InceptionV3 [21], a CNN trained on 1.2 million natural images across 1,000 classes; RadInceptionV3 [14], an InceptionV3 model trained on 1.35 million CT, MRI, and ultrasound images spanning neurological, oncological, and abdominal imaging; BiomedCLIP [26], a ViT trained on 15 million biomedical image-text pairs from 4.4 million scientific articles; MedDINOv3 [11], a ViT trained on 3.9 million 2D abdominal CT images; and MM-DINOv2 [19], a ViT trained on 50,000 brain MRI volumes. As shown in Table 3, the optimal representation is dataset dependent: MedDINOv3 achieves the best overall performance on the abdominal CT datasets, LiTS and MCT-LTDiag, whereas BiomedCLIP performs best on the BraTS brain MRI dataset. These results support the use of domain- and modality-appropriate feature extractors when applying FD-loss to medical image generation.

Distance Function To validate the choice of Fréchet Distance, we perform an ablation study substituting alternative distance functions. We investigate the impact of only tracking the first or second statistical moment, and using alternative distance functions for probability distributions such as symmetric KL divergence and Bhattacharyya distance. Calculating both symmetric KL divergence and Bhattacharyya distance require inverting Σ, which could introduce instability into training. To avoid this, we use Tikhonov regularization: $\Sigma_{stable} = \Sigma + \epsilon I$; we select $\epsilon = 10^{-6}$. We test each distance function with the LiTS dataset [3] in the MedDINOv3 representation space [11]. The results of the ablation study are reported in Table 4, which showed that utilizing FD yields the best downstream task performance.

5 Conclusion

In this work, we evaluated Fréchet Distance loss as a finetuning objective for segmentation-guided medical image generative models. Across abdominal CT and brain MRI datasets, FD-loss improved synthetic image realism and increased downstream segmentation performance, with the largest gains observed for heterogeneous tumor regions. Our encoder study showed that the representation space used for FD-loss is an important design choice, with medical and biomedical encoders generally outperforming natural-image features while varying in effectiveness across datasets and modalities. These results suggest that feature-distribution matching is a useful regularizer for generating synthetic data that better supports medical image segmentation. Future work should extend this framework to 3D image generation and other clinical tasks, including domain harmonization and image-to-image translation.

References

1. Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J.S., Freymann, J.B., Farahani, K., et al.: Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. Scientific Data **4**, 170117 (2017). https://doi.org/10.1038/sdata.2017.117
2. Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., Shinohara, R.T., Berger, C., et al.: Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge (2019). https://doi.org/10.48550/arXiv.1811.02629
3. Bilic, P., Christ, P., Li, H.B., Vorontsov, E., Ben-Cohen, A., Kaissis, G., Szeskin, A., Jacobs, C., et al.: The liver tumor segmentation benchmark (LiTS). Medical Image Analysis **84**, 102680 (2023). https://doi.org/10.1016/j.media.2022.102680
4. Bińkowski, M., Sutherland, D.J., Arbel, M., Gretton, A.: Demystifying MMD GANs (2021). https://doi.org/10.48550/arXiv.1801.01401
5. Cardoso, M.J., Li, W., Brown, R., Ma, N., Kerfoot, E., Wang, Y., Murrey, B., Myronenko, A., et al.: MONAI: An open-source framework for deep learning in healthcare (2022). https://doi.org/10.48550/arXiv.2211.02701
6. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a local nash equilibrium (2018). https://doi.org/10.48550/arXiv.1706.08500
7. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Advances in Neural Information Processing Systems. vol. 33, pp. 6840–6851 (2020), https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf
8. Jayasumana, S., Ramalingam, S., Veit, A., Glasner, D., Chakrabarti, A., Kumar, S.: Rethinking FID: Towards a better evaluation metric for image generation. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9307–9315. IEEE (2024). https://doi.org/10.1109/CVPR52733.2024.00889
9. Konz, N., Chen, Y., Dong, H., Mazurowski, M.A.: Anatomically-controllable medical image generation with segmentation-guided diffusion models (2024). https://doi.org/10.48550/arXiv.2402.05210

10. Konz, N., Osuala, R., Verma, P., Chen, Y., Gu, H., Dong, H., Chen, Y., Marshall, A., et al.: Fréchet radiomic distance (FRD): A versatile metric for comparing medical imaging datasets. Medical Image Analysis **110**, 103943 (2026). https://doi.org/10.1016/j.media.2026.103943
11. Li, Y., Wu, Y., Lai, Y., Hu, M., Yang, X.: MedDINOv3: How to adapt vision foundation models for medical image segmentation? (2025). https://doi.org/10.48550/arXiv.2509.02379
12. Mao, J., Wang, Y., Tang, Y., Xu, D., Wang, K., Yang, Y., Zhou, Z., Zhou, Y.: MedSegFactory: Text-guided generation of medical image-mask pairs (2025). https://doi.org/10.48550/arXiv.2504.06897
13. Mathiasen, A., Hvilshøj, F.: Backpropagating through fréchet inception distance (2021). https://doi.org/10.48550/arXiv.2009.14075
14. Mei, X., Liu, Z., Robson, P.M., Marinelli, B., Huang, M., Doshi, A., Jacobi, A., Cao, C., Link, K.E., et al.: RadImageNet: An open radiologic deep learning research dataset for effective transfer learning. Radiology: Artificial Intelligence **4**(5), e210315 (2022). https://doi.org/10.1148/ryai.210315
15. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., et al.: The multimodal brain tumor image segmentation benchmark (BRATS). IEEE transactions on medical imaging **34**(10), 1993–2024 (2015). https://doi.org/10.1109/TMI.2014.2377694
16. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation (2015). https://doi.org/10.48550/arXiv.1505.04597
17. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training GANs (2016). https://doi.org/10.48550/arXiv.1606.03498
18. Salimans, T., Ho, J.: Progressive distillation for fast sampling of diffusion models (2022). https://doi.org/10.48550/arXiv.2202.00512
19. Scholz, D., Erdur, A.C., Ehm, V., Meyer-Baese, A., Peeken, J.C., Rueckert, D., Wiestler, B.: MM-DINOv2: Adapting foundation models for multi-modal medical image analysis (2025). https://doi.org/10.48550/arXiv.2509.06617
20. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models (2022). https://doi.org/10.48550/arXiv.2010.02502
21. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2818–2826. IEEE (2016). https://doi.org/10.1109/CVPR.2016.308
22. Woodland, M., Castelo, A., Al Taie, M., Albuquerque Marques Silva, J., Eltaher, M., Mohn, F., Shieh, A., Kundu, S., et al.: Feature extraction for generative medical imaging evaluation: New evidence against an evolving trend. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. pp. 87–97. Springer Nature Switzerland (2024). https://doi.org/10.1007/978-3-031-72390-2_9
23. Wu, X., Su, H., Hua, Y., Xu, Y., Wang, L., Wang, X., Wang, S., Jin, B., et al.: A multi-phase CT dataset for automated differential diagnosis of liver tumors. Scientific Data **13**(1), 31 (2025). https://doi.org/10.1038/s41597-025-06343-4
24. Yang, J., Geng, Z., Ju, X., Tian, Y., Wang, Y.: Representation fréchet loss for visual generation. https://doi.org/10.48550/arXiv.2604.28190
25. Zhang, L., Jindal, B., Alaa, A., Weinreb, R., Wilson, D., Segal, E., Zou, J., Xie, P.: Generative AI enables medical image segmentation in ultra low-data regimes. Nature Communications **16**(1), 6486 (2025). https://doi.org/10.1038/s41467-025-61754-6

26. Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., et al.: BiomedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs (2025). https://doi.org/10.48550/arXiv.2303.00915