npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-024-01290-7

Privacy enhancing and generalizable deep learning with synthetic data for mediastinal neoplasm diagnosis

Check for updates

Zhanping Zhou[1,2], Yuchen Guo[2] ✉, Ruijie Tang[1,2], Hengrui Liang[3], Jianxing He[3] & Feng Xu[1,2] ✉

The success of deep learning (DL) relies heavily on training data from which DL models encapsulate information. Consequently, the development and deployment of DL models expose data to potential privacy breaches, which are particularly critical in data-sensitive contexts like medicine. We propose a new technique named DiffGuard that generates realistic and diverse synthetic medical images with annotations, even indistinguishable for experts, to replace real data for DL model training, which cuts off their direct connection and enhances privacy safety. We demonstrate that DiffGuard enhances privacy safety with much less data leakage and better resistance against privacy attacks on data and model. It also improves the accuracy and generalizability of DL models for segmentation and classification of mediastinal neoplasms in multi-center evaluation. We expect that our solution would enlighten the road to privacy-preserving DL for precision medicine, promote data and model sharing, and inspire more innovation on artificial-intelligence-generated-content technologies for medicine.

Deep learning (DL) has served as a fundamental technology and empowered many practical applications in medical image-based diagnosis and treatment[1–3]. The fascinating performance of DL models comes from the carefully designed algorithms, high-performance computing devices, and most fundamentally, the large-scale training data of high quality. Despite the promising performance and wide deployment of DL models in clinical applications, it is necessary to rethink a question: would these DL models leak the training images and more importantly, the privacy of numerous patients that provided these images? The answer is pessimistic because with the rapid development of privacy attack techniques, privacy leakage risks may be throughout the whole pipeline of DL-based medical image diagnosis, including data leakage in data inspection and data sharing, data memorization attacks during model training[4,5], data leakage from gradients during federated learning[6,7], white-box model inversion attacks[8–10] and black-box membership inference attacks[11–13] during and after model deployment (Fig. 1a). With these risks on medical images, two types of privacy may be violated: membership privacy[14] and identity privacy[15]. When membership privacy is violated, attackers may leverage membership information to infer sensitive attributes related to the data, such as health conditions and treatments. The violation of identity privacy is more serious, where attackers can directly recover the images and use them as identifiers of patients. This means that attacks can precisely recognize every single patient in the dataset and obtain their sensitive attributes. Although there are some attempts to address these issues during model training, including knowledge distillation[16,17], adversarial regularization[18,19], and differentially private training[20], it should be noted that these strategies either do not physically isolate real data from DL model or hinder model's performance due to the lack of real data. How to develop high performance DL models for medical image analysis in a privacy-preserving way becomes an urgent problem in the interdisciplinary field of artificial intelligence and medicine.

We argue that an inherent loophole in making privacy attacks work is the use of real images. In the widely used paradigm, the DL model is directly trained with the real images, making it feasible to obtain, reconstruct, or infer the real images. Therefore, we investigate if it is possible to train DL models solely on synthetic data to cut off the direct direction between models and real data using artificial-intelligence-generated-content (AIGC) techniques (Fig. 1b). In this way, privacy attacks can only get information of the synthetic images instead of real images, thus privacy leakages can be reduced. We name this scheme as data double, which is quite similar to the use of body doubles in movies. Despite its potential and simple principle, there are still three fundamental challenges associated with the data double: (1) How to generate synthetic images enabling diverse and complicated medical tasks, such as lesion segmentation, pathology subtype classification, and model interpretability analysis? (2) How to make synthetic-image-trained models superior, or at least comparable to real-image-trained models. (3) How about the effectiveness of data double against privacy

[1]School of Software, Tsinghua University, Beijing, China. [2]Beijing National Research Center for Information Science and Technology (BNRist), Tsinghua University, Beijing, China. [3]Department of Thoracic Oncology and Surgery, China State Key Laboratory of Respiratory Disease & National Clinical Research Center for Respiratory Disease, the First Affiliated Hospital of Guangzhou Medical University, Guangzhou, China. ✉e-mail: yuchen.w.guo@gmail.com; feng-xu@tsinghua.edu.cn

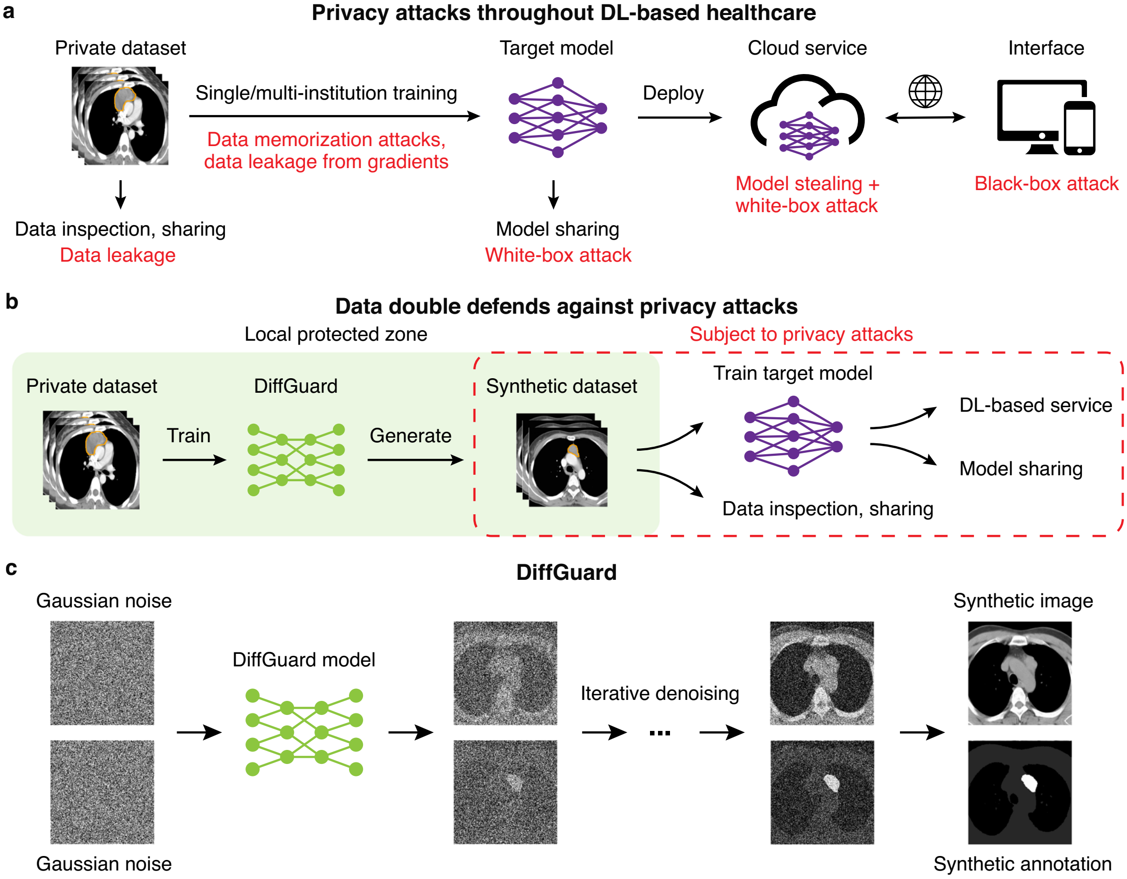

Fig. 1 | Data double defends against privacy attacks using DiffGuard model. **a** Deep learning applications for healthcare are subject to various privacy attacks throughout development and deployment. **b** Data double trains a generative model called DiffGuard (**c**) and uses it to generate a synthetic dataset in local, and downstream tasks use the synthetic dataset instead of the private dataset. **c** DiffGuard generates paired image and pixel-wise annotation by sampling Gaussian noise and iteratively denoising. We use resources from uxwing.com.

attacks? Although there are some attempts to lower privacy leakage by using synthetic medical images[21–23], the above problems have been rarely noticed or investigated. We believe providing effective and elaborate solutions for them will pave the way for privacy protection in DL-based medicine and the combination of AIGC and medicine.

We present DiffGuard, a data double technique for DL-based medical image analysis that supports multiple tasks, yields outstanding performance, and enhances privacy (Fig. 1c). Instead of adopting generative adversarial network (GAN) which is used by most medical image synthesis studies[21–23], we adopt the state-of-the-art diffusion model and extended it to simultaneously generate medical images and pixel-level annotations. A DiffGuard model can generate images and the corresponding annotations of multiple pathologies, supporting common tasks including pathology detection, lesion segmentation, and subtype classification. To improve the correctness of anatomical structures, we introduce an additional structure label mask to encode medical knowledge, so that the generative model pays attention to learning the shape of other anatomies and their relative position to lesions. To our best knowledge, DiffGuard is the first to demonstrate that models trained purely with synthetic data can achieve comparable or even higher performance than models trained with real data in medical image analysis applications. Additionally, it can enhance privacy safety against white-box and black-box attacks for data sharing and model deployment.

We evaluated the effectiveness of DiffGuard in the real-world medical application of mediastinal neoplasm diagnosis. Mediastinal neoplasms are thoracic diseases with all-encompassing pathological compartments that range from benign masses to malignant tumors[24–27]. Patients with malignant tumors may suffer from poor prognosis, thus accurate and timely diagnoses of mediastinal neoplasms are essential for better treatments and personalized healthcare[28]. In this study, we collected contrast-enhanced CT datasets and plain CT datasets for internal and external validation from 13 centers, and annotated five types of mediastinal neoplasms in the CT volumes. We trained DiffGuard on the axial CT slices and generated large-scale synthetic 2D CT images with corresponding pixel-level annotations. We first qualitatively assessed the synthetic images. Most of the synthetic CT images are of high quality and can hardly be visually distinguished from real images by human experts. We also trained an ablation model without using the structure label mask, and found that a proportion of synthetic images contain unrealistic anatomical structures. Then we quantitatively evaluated the utility of synthetic CT images generated by DiffGuard on two common diagnostic tasks, lesion segmentation and subtype classification. We designed a simple yet effective method to enhance the inter-slice continuity of 2D segmentation predictions. Experiment results demonstrated that DL models trained on DiffGuard-generated 2D CT images outperform the models trained on both 2D and 3D real data. Furthermore, we empirically evaluated the membership privacy of DiffGuard-generated images against a similarity-based threat model[29], and evaluated the membership privacy of trained DL models against a black-box membership inference attack method[30]. Compared with several baseline methods, DiffGuard successfully defends against privacy attacks while achieving the best model performance.

Table 1 | Characteristics of contrast-enhanced CT and plain CT datasets

Characteristic	Contrast-enhanced CT			Plain CT		
	Training set	Internal test set	External test set	Training set	Internal test set	External test set
Thymoma	276 (37.5%)	268 (37.4%)	56 (33.9%)	107 (35.8%)	174 (37.4%)	69 (37.9%)
Benign cysts	248 (33.7%)	237 (33.1%)	57 (34.5%)	115 (38.5%)	169 (36.3%)	59 (32.4%)
Thymic carcinoma	67 (9.1%)	65 (9.1%)	8 (4.8%)	19 (6.4%)	40 (8.6%)	12 (6.6%)
Germ cell tumor	53 (7.2%)	53 (7.4%)	17 (10.3%)	15 (5.0%)	22 (4.7%)	12 (6.6%)
Neurogenic tumor	92 (12.5%)	94 (13.1%)	27 (16.4%)	43 (14.4%)	60 (12.9%)	30 (16.5%)
Total patients	736	717	165	299	465	182
Age (years, mean ±s.d.)	50 ± 14	50 ± 14	51 ± 14	50 ± 13	49 ± 14	50 ± 14
Female	381 (51.8%)	374 (52.2%)	92 (55.8%)	155 (51.8%)	238 (51.2%)	83 (45.6%)

We believe that DiffGuard has the potential to be applied to other medical imaging modalities and tasks, and enlighten both research and application of computer-aided diagnosis.

Results

Datasets

We constructed a real-world multi-center dataset for mediastinal neoplasms diagnosis including lesion segmentation and subtype classification. The dataset had contrast-enhanced CT and plain CT for five types of mediastinal neoplasms, i.e., thymoma, benign cyst, thymic carcinoma, germ cell tumor, and neurogenic tumor. We filtered contrast-enhanced CT and plain CT images scanned between January 1st, 2010 and October 31st, 2020 from eight centers for model development and internal validation, and another five centers for external validation (Supplementary Table 1). The total thirteen medical centers were located in eight different provinces of China, spanning from north to south and from east to the middle-western region (Supplementary Figure no. 1). In each center, two board-certified pathologists including one senior pathologist examined the paraffin wax pathological sections and retained the cases that contained at least one of the five types of mediastinal neoplasms, where the gold standard pathology results were used as the ground truth classification label. From the internal centers, we collected 1453 contrast-enhanced CT scans and 764 plain CT scans, which were randomly split into the training dataset containing 736 contrast-enhanced CT scans and 299 plain CT scans, and the internal test set containing 717 contrast-enhanced CT scans and 465 plain CT scans. There was no patient overlapping between any training and test sets. From the external centers, we collected 165 contrast-enhanced CT scans and 182 plain CT scans with mediastinal neoplasms. The mediastinal neoplasm in each CT scan was delineated by one of the six board-certified radiologists individually using Materialise's interactive medical image control system (Materialise Mimics V20.0, Ghent, Belgium). The characteristics of the datasets are summarized in Table 1. We calculated the CT scanner statistics of contrast-enhanced CT scans and plain CT scans respectively (Supplementary Tables 2 and 3). There exist obvious distribution shifts between the internal training set, internal test set and external test set, which are proper to evaluate model's generalizability across CT scanners.

To further evaluate the cross-race generalizability of DiffGuard, we used data from the National Lung Screening Trial (NLST)[31] obtained from the National Cancer Data Access System of the US National Cancer Institute (NCI), through a data transfer agreement between the authors and the NCI (project number 868). Because there is no official diagnosis related to mediastinal neoplasms, we collected 11,162 participants with the official diagnosis of no significant abnormalities for all three screening times as the NLST test set.

Qualitative evaluation of DiffGuard-generated images

DiffGuard is capable of synthesizing high-quality medical images with precise annotations indistinguishable from real ones. For each CT modality, we trained two DiffGuard models: for generating images with and without mediastinal neoplasms respectively. For each mediastinal neoplasm subtype and normal control, we randomly gathered 20 real images and 20 DiffGuard-generated images (Fig. 2a, b), summing up to a total of 120 real images and 120 DiffGuard-generated images for each CT modality. More examples of synthetic images are shown in Supplementary Figs. 2 and 3. After, we shuffled the images, 3 radiologists with 7, 8, and 7 years of experience examined the images independently. They were informed that about half of the images were real and the other half were generated, and they were requested to infer whether each image was real or not. On contrast-enhanced CT, 56.7%, 50.0%, and 56.7% of DiffGuard-generated images were predicted to be real by the radiologists respectively, while only 43.3%, 43.3%, and 58.3% of real images were predicted to be real respectively (Supplementary Table 4). On plain CT, 53.3%, 56.7%, and 55.0% of DiffGuard-generated images were predicted to be real by the radiologists respectively, while only 46.7%, 43.3%, and 58.3% of real images were predicted to be real respectively (Supplementary Table 5). This indicates that the majority of DiffGuard-generated images are as realistic as real images.

To evaluate whether the proposed structure label mask enhanced the correctness of anatomical structures, we trained an ablation model without using the structure label mask, and examined the synthetic images. We found that a proportion of synthetic images contain unrealistic anatomical structures. As illustrated in Fig. 2c–g, the mediastinal neoplasms appear at wrong locations, and sometimes the shape of the trachea is unrealistic (Fig. 2d). In comparison, we rarely found such unrealistic anatomical structures in DiffGuard-generated images. This indicates that the structure label mask can help the generative model learn about correct anatomical structures.

Quantitative evaluation of DiffGuard-generated images

We adopted three commonly used metrics for generative model evaluation, i.e., Fréchet inception distance (FID)[32], kernel inception distance (KID)[33] and inception score (IS)[34]. FID and KID measure the distance between generated data distribution and real data distribution, where lower values indicate higher similarity with real data distribution. IS measures how realistic and diverse the generated images are, where a higher value indicates better quality and diversity. We compared DiffGuard-generated images with randomly augmented training images and AsynDGAN-generated images regarding contrast-enhanced CT and plain CT respectively (Supplementary Tables 4 and 5). It is not surprising that the FID and KID between training images and augmented training images are very low. The FID between DiffGuard-generated images and training images was 22.52 on contrast-enhanced CT and 35.20 on plain CT, higher than augmented training images (15.37 and 25.77) and significantly lower than AsynDGAN-generated images (40.01 and 53.42). Similarly, the KID between DiffGuard-generated images and training images was 0.017 on contrast-enhanced CT and 0.022 on plain CT, slightly higher than augmented training images (0.009 and 0.013) and significantly lower than AsynDGAN-generated images (0.030 and 0.036). These results indicate that DiffGuard learned training data distribution quite well, much better than AsynDGAN. Besides, DiffGuard achieved the highest IS among the three methods (5.80 and 5.90),

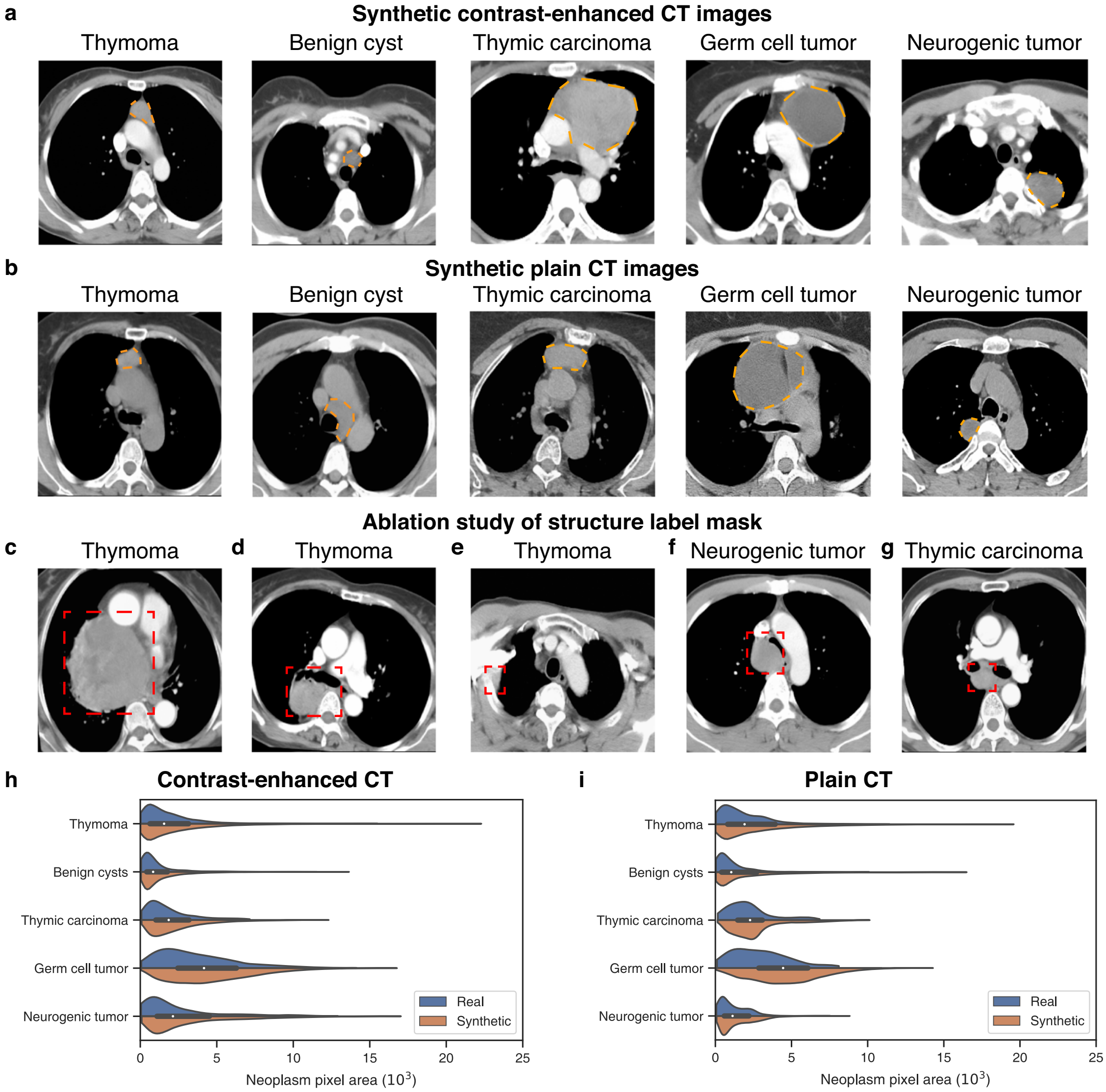

Fig. 2 | DiffGuard generates high quality images. a, **b** Examples of synthetic contrast-enhanced CT images (**a**) and plain CT images (**b**). **c**–**g** Unrealistic images generated by ablation method without structure label mask. **h**, **i** Mediastinal neoplasm size distribution in real and synthetic contrast-enhanced CT images (**h**) and plain CT images (**i**). DiffGuard generated mediastinal neoplasms of diverse sizes and covered real distribution well. Mediastinal neoplasms are delineated by dashed orange lines, and the unrealistic structures are annotated by red bounding boxes.

demonstrating its superior capability to generate diversified images. We further calculated the neoplasms size distribution (Fig. 2h, i), which also illustrates that DiffGuard learns the distribution of training data well and the synthetic neoplasms are diverse in size.

Evaluating utility of DiffGuard-generated images for mediastinal neoplasm segmentation

The utility of DiffGuard-generated images was first assessed on the mediastinal neoplasm segmentation task. We compared the performance of the model trained on the DiffGuard-generated images with the models trained on real CT images and synthetic images generated by AsynDGAN[35] which is a state-of-the-art GAN-based medical image generation method. For DiffGuard and AsynDGAN, we used 10,000 synthetic images for each type of mediastinal neoplasm. For real CT images, we used all our data as is elaborated in Table 1. Notice that the synthetic data is larger than the real one as we can synthesize as much as we wish. Three canonical types of neural network architectures were used: the widely overwhelming nnU-Net[36], the classical U-Net[37], and the recently popular TransUNet[38]. The model performance was evaluated using the Dice Similarity Coefficient (DSC) on the internal test set as well as the external test set, respectively.

On both the contrast-enhanced CT and plain CT datasets, the best segmentation DSCs were achieved by nnU-Net. With regard to contrast-enhanced CT, nnU-Net trained on DiffGuard-generated images achieved DSCs of 0.832(95% CI = 0.817-0.847) and 0.863(95% CI = 0.840–0.884) on the internal and external test sets respectively, over 20% higher than the DSCs of the models trained on the real CT images and AsynDGAN-generated images (Table 2, Supplementary Table 8). With regard to plain CT, nnU-Net trained on DiffGuard-generated images achieved DSCs of 0.795(95% CI = 0.774–0.816) and 0.781(95% CI = 0.745–0.815) on the internal and external test set respectively, over 40% higher than the DSCs of the models trained on the real CT images and AsynDGAN-generated images (Table 3, Supplementary Table 11). The superiority of DiffGuard was also validated on the U-Net and TransUNet models (Supplementary Tables 9,10,12,13).

The aforementioned models are all 2D models and then we compared the DiffGuard-trained 2D models with 3D models trained with real CT images. With regard to contrast-enhanced CT, the nnU-Net 2D model trained on the DiffGuard-generated images achieved comparable DSC to the nnU-Net 3D model on the internal test set and significantly outperformed the 3D model by a margin of 0.125 on the external test set. With

Table 2 | Evaluation of data utility on contrast-enhanced CT images. DSC: dice similarity coefficient

Training data	Internal test set			External test set		
	Segmentation DSC	Classification F1 score	Classification accuracy	Segmentation DSC	Classification F1 score	Classification accuracy
Real(2D)	0.695(0.673–0.716)	0.542(0.502–0.578)	0.647(0.614–0.682)	0.603(0.549–0.660)	0.496(0.400–0.583)	0.618(0.539–0.691)
Real(3D)	0.834(0.818–0.849)	0.663(0.618–0.705)	0.732(0.699–0.764)	0.738(0.682–0.786)	0.582(0.476-0.671)	0.624(0.552–0.703)
AsynDGAN	0.578(0.548–0.608)	0.491(0.444–0.531)	0.551(0.515–0.589)	0.604(0.548–0.670)	0.575(0.482–0.652)	0.612(0.539–0.685)
DiffGuard	0.832(0.816–0.847)	0.666(0.627–0.704)	0.714(0.679–0.749)	0.863(0.838–0.884)	0.640(0.548–0.729)	0.691(0.624–0.764)

Table 3 | Evaluation of data utility on plain CT images

Training data	Internal test set			External test set		
	Segmentation DSC	Classification F1 score	Classification accuracy	Segmentation DSC	Classification F1 score	Classification accuracy
Real(2D)	0.538(0.506–0.571)	0.487(0.424–0.539)	0.598(0.551–0.643)	0.471(0.420–0.524)	0.512(0.438–0.576)	0.593(0.522–0.665)
Real(3D)	0.760(0.733–0.787)	0.576(0.514–0.632)	0.656(0.613–0.697)	0.714(0.666–0.762)	0.574(0.486–0.649)	0.632(0.560–0.703)
AsynDGAN	0.550(0.514–0.586)	0.493(0.429–0.553)	0.570(0.527–0.613)	0.551(0.496–0.613)	0.409(0.328–0.478)	0.505(0.434–0.582)
DiffGuard	0.795(0.774–0.816)	0.596(0.536–0.652)	0.682(0.639–0.723)	0.781(0.746–0.815)	0.645(0.555–0.719)	0.709(0.643–0.775)

DSC: dice similarity coefficient.

regard to plain CT, the nnU-Net 2D model trained on the DiffGuard-generated images outperformed the nnU-Net 3D model by 0.035 on the internal test set and by 0.065 on the external test set. The experimental results demonstrate that 2D models trained on DiffGuard-generated data can match and even outperform both 2D and 3D models trained on real images in the mediastinal neoplasm segmentation task. Besides, the superiority is more significant on the external test set, demonstrating that training on large-scale DiffGuard-generated images can enhance model's generalizability. We show and compare some test predictions by the nnU-Net models trained on real CT images (2D and 3D version), AsynDGAN-generated images and DiffGuard-generated images in Supplementary Figs. 4 and 5.

We also evaluated the relationship between the model performance and the size of mediastinal neoplasm. For all the mediastinal neoplasms in the internal and external test sets, we sorted their volumes in ascending order and uniformly divided them into five groups. In each group, we compared the DSCs of the nnU-Net models trained on the real CT images (2D and 3D version), AsynDGAN-generated images, and DiffGuard-generated images. On both the contrast-enhanced datasets (Supplementary Fig. 6a) and plain CT datasets (Supplementary Fig. 6b), the DSC of bigger mediastinal neoplasms was higher than the smaller ones. In all the experiments, the DiffGuard-trained models achieved comparable and sometimes better performance than the models trained on other data, and the variances of their DSCs were much smaller than other models. The superiority of DiffGuard-trained models over the other models was most obvious in small mediastinal neoplasms, which were more difficult to recognize and delineate and thus more important in medical practice. The above experimental results indicate that large-scale DiffGuard-generated images can enhance the performance and robustness of mediastinal neoplasm segmentation models, particularly on the hard problem of small neoplasm segmentation.

Some previous studies showed 3D models outperformed 2D models when trained on the same CT volumes. One reason is that 3D models can leverage the inherent continuity of CT slices to perceive the shape and texture in 3D space, while 2D models perceive the CT slices separately. Nonetheless, DiffGuard can generate large-scale and diverse synthetic images to train more accurate and robust 2D models than the models trained on limited real images. Besides, we designed a simple yet effective post-processing method for 2D models to leverage the principle of continuity, which significantly improved the segmentation DSC on contrast-enhanced CT and plain CT (Supplementary Tables 14,15). In Supplementary Fig. 7, we show the scan predictions by the nnU-Net trained on DiffGuard-generated images without and with the post-processing method. Taking advantage of the adjacent slices, the post-processing method successfully eliminates the false positive predictions. From the above results, it seems that it is not necessary to use 3D volumes for accurate diagnosis. With 2D images and simple post-processing step incorporating continuity, the performance could be even better, especially for external test. In addition, it is more efficient and easier to train a 2D generative model and generate much more images for training, which can further improve the performance. We believe these interesting findings could inspire further studies in AIGC for medicine.

We further evaluated the cross-population generalizability of DiffGuard on CT scans from NLST. Because there is no available examination result on mediastinal neoplasms, we used the model trained with DiffGuard to filter out some suspected mediastinal neoplasms. Three radiologists including two senior radiologists checked these cases, among which they identify and delineated the thymoma in case 206366 and and the benign cyst in case 217293. On the two cases, we compared the segmentation predictions of the models trained with different strategies (Supplementary Fig. 8). All models except the model trained on 2D read images successfully detected the neoplasms. Models trained on AsynDGAN-generated images and DiffGuard-generated images achieved more accurate segmentation prediction than models trained on real images. The experimental results demonstrate that DiffGuard improves model generalizability across populations.

Evaluating utility of DiffGuard-generated images for mediastinal neoplasm classification

The type of mediastinal neoplasm is classified on the basis of segmentation prediction, which is described in the subsection 'Mediastinal neoplasm segmentation and classification' of the Method section. The classification performance was evaluated using two metrics: macro-average F1 score and accuracy score. With regard to contrast-enhanced CT, nnU-Net trained on DiffGuard-generated images achieved a F1 score of 0.666(95% CI = 0.628–0.701) and an accuracy score of 0.714(95% CI = 0.682–0.745) on the internal test set, and achieved a F1 score of 0.640(95% CI = 0.544–0.726) and an accuracy score of 0.691(95% CI = 0.624–0.758) on the external test set (Table 2, Supplementary Table 8). It significantly outperformed the 2D nnU-Net models trained on the real CT images and AsynDGAN-

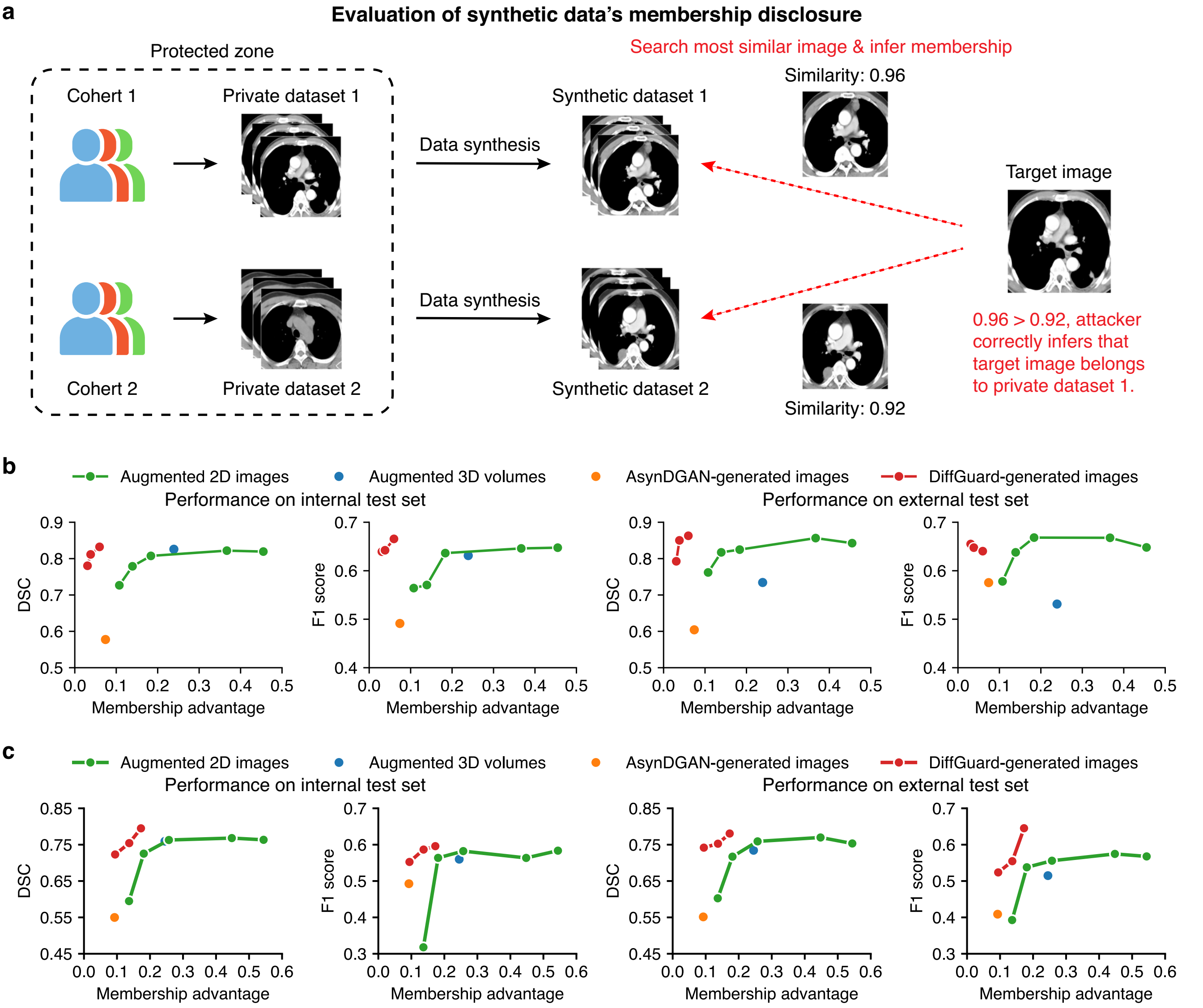

Fig. 3 | Empirical evaluation of membership privacy disclosure from DiffGuard-generated images. a For a real image, the attacker searched the most similar synthetic images and infer whether the corresponding patient was included in the training cohort. **b**, **c** Comparison of the data utility and privacy safety on contrast-enhanced CT (**b**) and plain CT (**c**). DiffGuard-generated images significantly lowered the membership inference advantage while maintaining high data utility.We use resources from uxwing.com and www.iconpacks.net.

generated images, with a F1 score margin of over 20% and an accuracy margin of over 10%. Compared with the 3D model, its F1 score and accuracy score were comparable on the internal test set, and were higher than the 3D model's scores by more than 10% on the external test set, demonstrating better generalizability. DiffGuard's superiority was even more significant on the plain CT datasets, where nnU-Net trained on DiffGuard-generated images achieved a F1 score of 0.596(95% CI = 0.535-0.655) and an accuracy score of 0.682(95% CI = 0.639-0.725) on the internal test set, and achieved a F1 score of 0.645(95% CI = 0.558–0.726) and an accuracy score of 0.709 (95% CI = 0.643–0.775) on the external test set, over 20% higher than the 2D baselines and both higher than the 3D model (Table 3, Supplementary Table 11). The superiority of DiffGuard was also validated on the U-Net and TransUNet models (Supplementary Tables 9,10,12,13). Ablation studies of the post-processing method show that F1 score and accuracy are improved in most of the experiments (Supplementary Tables 14,15). The aforementioned experimental results demonstrate that models trained on DiffGuard-generated images match and even outperform models trained on real images in the classification task, especially for external test.

Influence of scaling up DiffGuard-generated images on model performance

We evaluated how scaling up DiffGuard-generated images influences the model performance. We increased the number of synthetic images from 500 to 10,000 and trained the nnU-Net, U-Net and TransUNet models. The experimental results of nnU-Net models are shown in Supplementary Figs. 9 and 10, and the detailed results of the above models are shown in Supplementary Tables 8-13. On both contrast-enhanced CT and plain CT, there is an obvious trend that the segmentation DSC, classification F1 scores and accuracy scores increase when more synthetic samples are used, and the marginal benefit of using more training data gradually diminishes. We believe DiffGuard learns the distribution of the training data well and with sufficient sampling in the distribution, the models for the downstream tasks can lean well form the distribution, while over sampling is not necessary as this cannot bring more information.

Empirical evaluation of membership privacy disclosure from DiffGuard-generated images

The defense effectiveness of data double against privacy attacks lies in the DiffGuard-generated images. On the one hand, the fact that synthetic images may be available to attackers requires them to contain as little privacy as possible. On the other hand, the less privacy they contain, the less privacy downstream DL models leak.

We empirically evaluate the membership privacy of DiffGuard-generated images. In terms of data leakage, membership privacy is much easier for attackers to violate than identity privacy, thus low membership privacy ensures low identity privacy. We followed the standard experimental setting of membership privacy assessment that evaluated how successfully attackers can leverage synthetic images to distinguish between private cohorts (Fig. 3a). Specifically, the synthetic dataset 1 and 2 were generated from the private dataset 1 and 2 using the same data synthesis method such as DiffGuard. Then the two private datasets were merged and randomly shuffled. With the two synthetic datasets, for each sample in the merged dataset, attackers attempt to infer whether it belongs to the private dataset 1 or 2. In this study, we used the internal training set and internal test set collected at the same medical centers as private dataset 1 and private dataset 2. We used the feature similarity between images to find potential copies. Specifically, for each target image, the attacker calculates the similarity scores between it and every image in the two synthetic datasets, and picks out the synthetic image with the highest score. If the most similar image to the target image belongs to the synthetic dataset 1, then the attack infers that it might occur in the private dataset 1, and vice versa. Privacy risk was measured using attack membership advantage[30] and attack accuracy. Membership advantage is defined as the difference between the attack recall and false positive rate, which generally ranges between 0 and 1. Higher membership advantage and attack accuracy indicate more membership disclosure. Generally, there is a trade-off between privacy and utility (discussed in previous subsections). Therefore, we compared the utility of synthetic images together with privacy safety.

We compared DiffGuard with three data synthesis baselines: (1) augmented 2D images, which randomly rotated and cropped the real CT images. We experimented on 250, 500, 1000, 6000 and 10,000 augmented images per class. (2) augmented 3D volumes, which randomly rotated and cropped the real CT volumes. We augmented all the CT volumes in the real dataset once. (3) AsynDGAN-generated images. We trained AsynDGAN models on the training dataset, and generated 10,000 images per class. To evaluate data utility, we used the nnU-Net 2D model for 2D images and the nnU-Net 3D model for 3D volumes.

Figure 3b shows the experimental results on contrast-enhanced CT. There is a clear trade-off between data utility and privacy safety for the other methods: the DSC and F1 score increase while the membership advantage also increases with more synthetic images. With 500 DiffGuard-generated images per class, the membership advantage is only 0.031(95% CI = 0.024–0.038) and the attack accuracy is only 0.515(95% CI = 0.512–0.519) in the contrast-enhanced CT experiment. When the number of images increases to 10,000 per class, there is a large improvement in DSC, while the membership advantage slightly increases to 0.060(95% CI = 0.053–0.067) and the attack accuracy slightly increases to 0.530(95% CI = 0.526–0.533).

Figure 3c shows the experimental results on plain CT. With 500 DiffGuard-generated images per class, the membership advantage is only 0.094(95% CI = 0.085–0.104) and the attack accuracy is only 0.550(95% CI = 0.545–0.555). When the number of images increases to 10,000 per class, there is a large improvement in both DSC and F1 score, while the membership advantage slightly increases to 0.173(95% CI = 0.164–0.182) and the attack accuracy increases to 0.601(95% CI = 0.597–0.606).

In comparison, the augmented 2D images and 3D volumes have much higher privacy risks when the data utility is comparable or inferior to DiffGuard-generated images. AsynDGAN-generated images have low privacy risks, but their data utility is the worst among all, indicating its low synthesis quality. Detailed experimental results are shown in Supplementary Tables 16,17. To sum up, DiffGuard achieved the best utility-privacy tradeoff on the mediastinal neoplasm datasets, and scaling up DiffGuard-generated images improves model performance with minimal privacy risks.

Empirical evaluation of privacy disclosure from DL models against black box attacks

In addition to potential privacy disclosure associated with data leakage, we also evaluated how DiffGuard could lower the privacy disclosure of DL models. Since model inversion attacks violate privacy by reconstructing training images, their privacy risks depend on the privacy risks of the training images, which have been evaluated in the former subsection. Here we focus on membership inference attacks that violate membership privacy by inferring whether a specific image is used to train a specific DL model. We consider a practical black-box setting where attackers can only obtain the model predictions of input images (Fig. 4a). Given models trained on the training dataset, we empirically evaluated how successful attackers can infer the membership of images in the training dataset and internal test set. We used the classic black-box membership inference attack method based on the generalization error[30], which is not dependent on data, model architecture, or task.

We compared DiffGuard with four groups of methods: (1) Custom training where models are trained on real CT images, including 2D and 3D versions. (2) Training on augmented images, including the augmented 2D images and the augmented 3D volumes mentioned in the previous section. (3) Training on real images with differential privacy. We used the most used DP-SGD method[20]. (4) Training on the AsynDGAN-generated images. As usual, we evaluated both the privacy safety and model utility. To evaluate model utility, we used the nnU-Net 2D model for 2D images and the nnU-Net 3D model for 3D volumes.

On contrast-enhanced CT (Fig. 4b), the attack against the model trained on the real 2D images achieved a membership advantage of 0.563(95% CI = 0.520–0.606) and an attack accuracy of 0.782(95% CI = 0.760–0.802), indicating severe privacy risks. The attack against the model trained on the real 3D volumes achieved a membership advantage of 0.176(95% CI = 0.123–0.229) and an attack accuracy of 0.588(95% CI = 0.562–0.615). Training on the DiffGuard-generated 10,000 images, the membership advantage is reduced to 0.091(95% CI = 0.040–0.143), and the attack accuracy is reduced to 0.545(95% CI = 0.520–0.571) while the model performance is comparable on the internal test set and even higher on the external test set.

On plain CT (Fig. 4c), the attack against the model trained on the real 2D images achieved a membership advantage of 0.830(95% CI = 0.786–0.869) and an attack accuracy of 0.911(95% CI = 0.889–0.931), indicating severe privacy risks. The attack against the model trained on the real 3D volumes achieved a membership advantage of 0.428(95% CI = 0.362-0.492) and an attack accuracy of 0.702(95% CI = 0.668–0.734). Training on the DiffGuard-generated 10,000 images, the membership advantage is reduced to 0.190(95% CI = 0.116–0.260) and the attack accuracy is reduced to 0.602(95% CI = 0.568–0.636) while the model performance is comparable on the internal test set and even higher on the external test set.

Compared with DiffGuard, other comparative methods have inferior performance-privacy trade-offs. Models trained on AsynDGAN-generated images or DP-SGD could reduce the membership privacy disclosure to a low level at the expense of a significant performance drop. Models trained on the augmented 2D images can achieve comparable performance when the number of images increases to 6,000 per class, but the privacy disclosure also rises rapidly. Models trained on the augmented 3D volumes slightly reduced the privacy disclosure at the expense of a slight performance drop. Detailed experimental results are shown in the Supplementary Tables 18,19.

Discussion

Previous privacy-enhancing methods for DL can be categorized into three groups. The first group is specifically designed for model training, including adversarial regularization[18,19] and knowledge distillation[16,17]. These methods cannot enhance privacy against attacks in data inspection and data sharing,

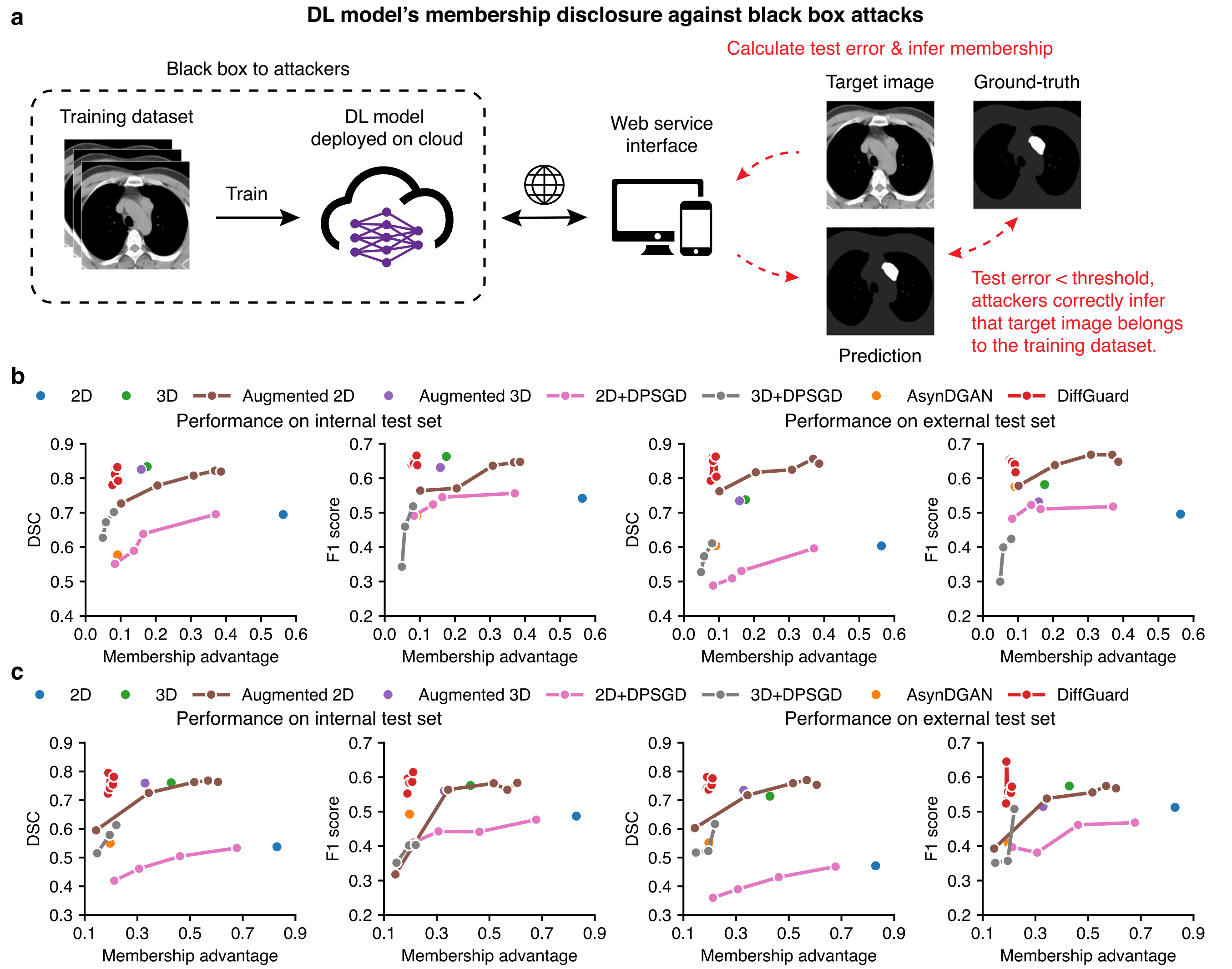

Fig. 4 | Defense effectiveness of the data double against black box membership inference attacks on DL models. a For a real image, the attacker obtained model's prediction, calculated the test error and infer whether the corresponding patient was included in the training cohort. **b**, **c** Comparison of the utility-privacy trade-off on contrast-enhanced CT (**b**) and plain CT (**c**). DiffGuard significantly lowered the membership inference advantage of victim models while maintaining high model utility. We use resources from uxwing.com.

data memorization attacks, and data leakage from gradients. Another major weakness is that models trained with these methods witness an obvious performance drop. In addition, knowledge distillation requires additional training images.

The second group of methods is based on differential privacy (DP), which was initially designed to anonymize the outputs of interactive queries to a database[39]. It can provide a strong privacy guarantee through theoretical analysis[39,40]. With the rise of DL techniques, there have been advances in training DL models with differential privacy[20]. Although DP provides a theoretical upper bound of privacy risk, it has been proved that the upper bound is often trivial and provides no effective privacy protection[30]. Another issue of DP is that the privacy preserving effectiveness is highly dependent on the context, making it hard to decide appropriate privacy budgets for different applications. Thus, empirical assessments should be used to evaluate the membership privacy[41]. Studies found that while DP methods could reduce models' privacy risks, they suffered from challenging hyperparameter tuning, significant model utility drop, and extremely long training time[42,43]. Recent studies on the use of DP in medical image analysis indicate that the utility decrease is more significant on smaller training datasets and more complicated tasks[44,45]. These findings are consistent with our experimental results that DP-trained models exhibit notably lower utility compared to those trained with DiffGuard, given that mediastinal neoplasm diagnosis has limited training data and features complicated segmentation and subtype classification tasks.

The last group of methods uses synthetic data instead of real data to lower privacy risks, which was first proposed for handling electronic health records (EHR) which contain sensitive information of patients[46–49]. With the rise of powerful deep generative models, particularly generative adversarial networks (GAN)[50], medical image synthesis has become increasingly popular for data augmentation[51], solving class-imbalance problem[52], and privacy-enhancing data sharing[21–23]. However, GANs suffer from training instability due to the Nash equilibrium[53], internal covariate shift[54], mode collapse[55,56], vanishing gradient[55], and lack of proper evaluation metrics[57]. As an alternative method, diffusion models beat GANs and achieve state-of-the-art performance on image synthesis[58,59], and synthetic images have been used to improve model generalizability[60,61]. Recently, it has been indicated that models trained purely with synthetic data can achieve comparable performance to the models trained with real data on natural image datasets[62,63]. With regard to medical images, previous studies have made the early attempt to leverage diffusion models for unconditional generation[64–70], class-conditional generation[71–75], parameter-conditional generation[76,77], text-conditional generation[78–81], image-conditional generation[82,83], generation conditioned by ground truth pixel-wise annotation[84–93], and co-generation of paired image and pixel-wise annotation[94–100]. Among them, the latter two kinds of methods provide corresponding pixel-wise annotation, which is the most fine-grained type of annotation and supports most medical image applications. However, these previous studies have three key limitations. Firstly, the utility of their synthetic images could not match the real images due to limited authenticity and diversity. Secondly, these studies did not quantitatively evaluate the privacy risks of synthetic images comprehensively. They did not consider the privacy risks for data sharing. Some even performed no privacy analysis, and some only qualitatively analyzed the most similar real image to a few synthetic images. Lastly, they did not evaluate the privacy safety of models trained on synthetic medical images against privacy attacks.

In this study, we propose DiffGuard that can generate images of multiple pathologies with pixel-wise annotations, meeting the needs of most downstream applications, such as segmentation, detection, and classification. We qualitatively and quantitatively evaluate the utility of DiffGuard-generated images for mediastinal neoplasm segmentation and classification tasks, which achieve even better model performance than real images. Furthermore, we evaluate the membership privacy disclosure of DiffGuard-generated images and DL models trained on them. The experiments indicate that DiffGuard significantly lowers privacy disclosure while maintains superb model generalizability.

DiffGuard is established by some key technical improvements. First, DiffGuard adopts a more advanced data synthesis scheme named denoising diffusion probabilistic model[58], which can generate images of higher quality and diversity than previous works[59]. This has been demonstrated in the comparison between AsynDGAN and DiffGuard. We extended the ability of the denoising diffusion probabilistic model to simultaneously generate images and annotations. This has two advantages: (1) the automatic generation of fine-grained annotations saves huge human annotation efforts. (2) co-training of the image and annotation aids model's understanding of the semantic structures, leading to higher image quality and annotation accuracy. In addition, we introduced an additional structure label mask to encode medical knowledge, so that the generative model paid attention to learn the shape and texture of all structures. We show that the synthetic images with structure label mask have more realistic structures than those without the structure label mask. The above technical improvements contribute to the success of DiffGuard.

We also investigate whether a 2D or 3D model is better for CT-based AI. 3D models have been the state-of-the-art method for a variety of segmentation tasks for a long time[36]. One possible reason is that 3D models can leverage the inherent continuity of CT slices to perceive the shape and texture in 3D space. However, recent studies show that 2D models can sometimes match and even outperform 3D models[101,102]. In this study, instead of training 3D DiffGuard models to synthesize 3D CT volumes, we designed a cost-effective post-processing method for 2D models to leverage the principle of continuity, which requires much less GPU memory and much less computation. Experimental results validated the effectiveness of the post-processing method in both segmentation and classification tasks. What this can inspire us is that in the age of deep learning and large models, fully exploiting the domain knowledge and task-related characteristics can achieve better performance with smaller models.

In the practical application of DiffGuard, it is important to balance utility, privacy, and resources by determining an appropriate synthetic dataset size. Generally, the number of synthetic images is correlated with a DL model's utility, the amount of computation, and the disk storage for storing the images. As the experimental results indicated, based on DiffGuard, we can easily achieve a cost-effective, accurate and privacy-preserving strategy by gradually increasing the number of synthetic images until the performance increment converges.

One limitation of our work is that DiffGuard does not provide provable privacy protection, i.e., guaranteeing an upper bound of privacy leakage just like differential privacy. Though we empirically demonstrate the effectiveness of DiffGuard against some currently representative privacy attack methods, how well would DiffGuard defend against future attack methods remains unknown. Therefore, the adoption of DiffGuard should be carefully examined in practical settings.

While DiffGuard alone does not offer provable privacy protection, we devised a compound scheme that combines DiffGuard with differential privacy. This scheme harnesses the complementary advantages of both methods, simultaneously achieving high model utility and theoretically provable privacy protection. To evaluate this compound strategy, we trained models on DiffGuard-generated images using the DPSGD method at varying privacy budgets. Compared with DiffGuard, it offers a theoretical privacy guarantee and the flexibility to adjust privacy budgets. Compared with DPSGD, it demonstrates superior trade-offs between model utility and privacy, ensuring better utility at equivalent levels of privacy safety (Fig. 5, Supplementary Tables 18,19).

Besides, there remain challenges and opportunities to further reduce the resources used by DiffGuard. As mentioned above, the principle of DiffGuard is based on the denoising diffusion probabilistic model[58], which can generate data of higher fidelity and diversity at the expense of more computational cost and carbon emission. Training DiffGuard took about 14 GPU days on NVIDIA RTX 3090 cards, and sampling 60,000 images took about 6.5 GPU days. Fortunately, there has been increasing effort to reduce the computational cost of training deep neural networks[103] and sampling diffusion models[104–106]. We employed the mixed precision training strategy[103], which significantly reduces GPU memory costs and communication burdens between GPUs, thus diminishing computation while maintaining model performance. We found the mixed precision training techniques reduced the training time by 27% to about 10 GPU days. To accelerate the data sampling process, we adopted a fast sampling method called DDIM[104]. Originally, DiffGuard generated an image by denoising a noisy image iteratively for 1000 steps, which leads to low generation speed. DDIM accelerates the sampling by reducing the denoising steps to as few as 100 without compromising the quality of synthetic data, which further reduces the sampling time by 90% to 0.65 GPU days.

Another limitation of our study regards the test datasets. We mainly experimented on data collected from hospitals in China. Since there is no publicly available datasets with ground truth from other populations, we could only filter small number of scans with potential mediastinal neoplasms from NLST. Future evaluation should include more data with ground truth labels from other populations.

Data double methods such as DiffGuard may have a broad impact on the industry and research field. Synthetic images may serve as data double and replace the role of real images for downstream applications, thus enhancing privacy safety. This may largely promote data sharing, leading to better collaboration between institutions and enhancing the reproducibility of research. Besides, the flexibility of data generation may help mitigate the bias of training data. Creating more diversified data for the minority group and balancing the sample number between groups may enhance the fairness of downstream models. Moreover, data double can also be used for creating digital twins of patients and their trajectories, which can be used to optimize treatment plans and improve patient outcomes.

While it is promising to adopt synthetic data for enhancing generalizability, mitigating bias and preserving privacy, it should be carefully examined in practical use. At present, there is no clear legislation surrounding the use of synthetic data[107], and current data protection regulations such as the General Data Protection Regulation (GDPR) and the Health Insurance Portability and Accountability Act (HIPAA) are limited in their ability to address all the potential risks associated with synthetic data[108,109]. This loophole could potentially be exploited by malicious entities, and the misuse of synthetic data may cause severe consequences. As we indicated in this research, synthetic data generated by ill-designed methods such as random augmentation still contain much patient information and may cause privacy leakage. Another issue is about potential unfairness caused by synthetic data. Synthetic data can also exacerbate data disparity

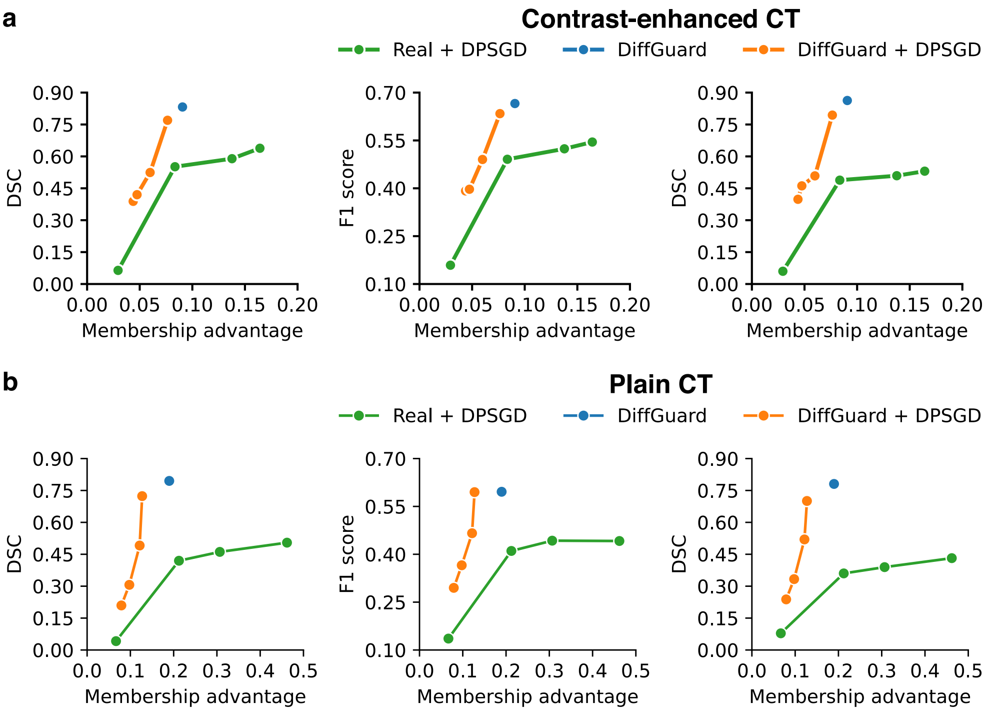

Fig. 5 | The compound scheme of DiffGuard and differential privacy. On both contrast-enhanced CT (**a**) and plain CT (**b**), this scheme achieves better privacy-utility trade-offs than DPSGD.

when digitally disadvantaged groups are not considered in method design[110]. These issues have been recognized and there have been calls to establish a digital chain-of-custody to fully supervise synthetic data throughout their life cycle[109,111]. We hope that this research can raise more attention to the use and misuse of synthetic data and contribute to a better way of using synthetic data for good.

Methods

Ethical Approval

This study was approved by the Ethics Committee of the National Center for Respiratory Medicine/The First Affiliated Hospital of Guangzhou Medical University (Date October 12th, 2020; IRB number: 2020 NO.138), Shanghai Chest Hospital (approved in year 2021), The First Affiliated Hospital, School of Medicine, Zhejiang University (approved in year 2021), Zhongshan City People's Hospital (approved in year 2021), Peking University Shenzhen Hospital (approved in year 2021), Fujian Medical University Union Hospital (approved in year 2021), Affiliated Cancer Hospital & Institute of Guangzhou Medical University (approved in year 2021), Sichuan Cancer Hospital & Institute (approved in year 2021), The Fourth Affiliated Hospital of China Medical University (approved in year 2021), The First Affiliated Hospital of Guangzhou Medical University, Gaozhou People's Hospital (approved in year 2021), Qingdao Municipal Hospital (approved in year 2021), and The First Affiliated Hospital of Xi'an Jiaotong University (approved in year 2021), respectively. Written informed consent to participate was obtained from all institutions.

Dataset collection

The collection of CT scans with mediastinal neoplasms followed the ITMIG's standard[112]. Each medical center retrospectively searched in their radiology database for chest CT scan interpretations that were performed between January 1st, 2010 and December 31st, 2021 and included either one of the following terms (in Chinese): "mediastinal nodule", "mediastinal lesion", "mediastinal neoplasm", or "mediastinal mass". Among the search results, we reviewed the chest CT reports and images to filter the CT scans by four criteria: (1) date of initial imaging showing the mediastinal abnormality; (2) imaging modality that showed the mediastinal lesion; (3) mediastinal compartment where the epicenter of the abnormality resided; (4) five types of pathology diagnosis according to WHO 2015[113] version of the mediastinal tumor from surgery or biopsy. We excluded recurrences or duplication. The study population included all patients of each age, and their age and sex were also collected.

Data preprocessing

Each CT scan was cropped around the lungs to remove useless margins, and sliced along the axial direction into images. Then the pixel values of the images were normalized to 0–1 using the window level at 0 and the window width at 400. The images and masks were resized to different sizes depending on the neural network architecture.

DiffGuard

DiffGuard can be applied to any medical imaging modality for paired image and annotation generation. Without loss of generality, we assume that there are C classes in the annotations, e.g., background, tissues, and pathologies, which are sequentially numbered $0, 1, \cdots, \mathrm{C}-1$. For an image y_{im}, its annotations are represented as a mask y_{mask} of the same size in which each pixel is assigned the number of the class that it is annotated. The image and mask are linearly normalized into the range between -1 and 1. For tasks where the annotation is needed, they are concatenated into a 3-dimensional array y. Otherwise, the normalized image is used without concatenation.

Data generation of DiffGuard is based on the diffusion models, which comprise a forward diffusion process and a reverse denoising process that is used at generation time. The forward diffusion process is a Markovian process that iteratively adds Gaussian noise to the sample $y_0 = y$ over T iterations:

$$p(y_{t+1}|y_t)=\mathcal{N}(y_{t-1};\sqrt{\alpha_t}y_{t-1},(1-\alpha_t)I) \tag{1}$$

$$p(y_1, y_2, \cdots, y_T|y_0) = \prod_{t=1}^{T} p(y_t|y_{t-1}) \tag{2}$$

where $\{\alpha_t\}_{t=1}^{T}$ are the hyper-parameters of the noise schedule. From Eq. (2), we can marginalize the forward process:

$$p(y_t|y_0)=\mathcal{N}(y_t;\sqrt{\gamma_t}y_0,(1-\gamma_t)I),t=1,2,\cdots,T \tag{3}$$

where $\gamma_t=\prod_{s=0}^{t}(1-\alpha_s)$. The above equations lay the foundation for the forward diffusion process. Our goal is to recover the target image y_0 given a noisy image y_T:

$$y_T=\sqrt{\gamma_T}y_0+\sqrt{1-\gamma_T}\epsilon,\epsilon\sim(0,I) \tag{4}$$

Therefore, if we can estimate ϵ, then we can calculate y_0:

$$y_0=\frac{1}{\sqrt{\gamma_t}}(y_t-\sqrt{1-\gamma_t}\epsilon) \tag{5}$$

To estimate ϵ, we train a neural network f parameterized by θ with the following objective:

$$\mathbb{E}_y\mathbb{E}_{(\epsilon,\gamma)}||f_\theta(y,\gamma)-\epsilon||_2^2 \tag{6}$$

We adopted the same network architecture and hyper-parameters as the method proposed by Ho et al.[58] and trained with AdamW optimizer[114] at a learning rate of 0.0001 for 1,000,000 iterations of batch size 12.

During the inference stage, instead of directly estimating y_0 using f, we iteratively perform the denoising for T steps. The posterior distribution of y_{t-1} given y_0 and y_t can be formulated as:

$$p(y_{t-1}|y_0,y_t)=\mathcal{N}(y_{t-1}|\mu_t,{\sigma_t}^2I) \tag{7}$$

where $\mu_t=\frac{\sqrt{\gamma_{t-1}}(1-\alpha_t)}{1-\gamma_t}y_0+\frac{\sqrt{\alpha_t}(1-\gamma_{t-1})}{1-\gamma_t}y_t$ and $\sigma_t=\sqrt{\frac{(1-\gamma_{t-1})(1-\alpha_t)}{1-\gamma_t}}$. Given Eq. (5) and Eq. (7), we can obtain the estimation:

$$\hat{\mu}_t=\frac{1}{\sqrt{\alpha_t}}(y_t-\frac{1-\alpha_t}{\sqrt{1-\gamma_t}}f_\theta(y_t,\gamma_t)) \tag{8}$$

Following Ho et al.[58], we use $\hat{\sigma}_t=\sqrt{1-\alpha_t}$. As a result, each iteration of the denoising process can be computed as:

$$\widehat{y_{t-1}}=\frac{1}{\sqrt{\alpha_t}}\left(y_t-\frac{1-\alpha_t}{\sqrt{1-\gamma_t}}f_\theta(y_t,\gamma_t)\right)+\sqrt{1-\alpha_t}\epsilon_t,t=T,T-1,\cdots,1 \tag{9}$$

where ϵ_t is randomly sampled from $\mathcal{N}(0,I)$. In practice, we set $T=1000$.

Finally, each sampled $\hat{y}$ is decoded into a synthetic image and its corresponding annotation mask. The former channels of $\hat{y}$ represents the synthetic image and are linearly rescaled to the range $(0,1)$. The last channel of $\hat{y}$ represents the annotation mask and is linearly rescaled to the range $(0,C)$. For each pixel, the floating-point value is converted to the class label with the smallest absolute difference.

To ensure the quality of synthetic datasets, we filtered out low-quality synthetic images with an autoencoder model. We also filtered out the synthetic samples with more than one type of mediastinal neoplasm.

Baseline methods

In this study, we compared DiffGuard with several baseline methods. For the 2D augmentation method, each axial CT slice in the real CT scans was randomly rotated by an angle uniformly sampled from −2° to 2°, then randomly cropped and resized to a height and weight of 256. For the 3D augmentation method, each CT scan was randomly rotated by an angle uniformly sampled from −5° to 5°, then randomly resized in three dimensions whose scales were sampled independently between 0.9 and 1.1. To make a fair comparison with AsynDGAN, we followed the original training hyperparameters except for the multi-institutional federated learning setting, and we trained the models on the gathered training images. We used the original training hyperparameters and adopted the same data augmentation strategies as DiffGuard, i.e., random horizontal flipping, rotation, and resizing. After training, we generated images with randomly rotated and resized ground-truth label masks. To evaluate the privacy-utility trade-off of DP-SGD, we set the noise multiplier as 0.01, 0.03, 0.05, 0.07 for 2D nnU-Net models, and 0.003, 0.005, 0.007 for 3D nnU-Net models.

Mediastinal neoplasm segmentation and classification

We trained segmentation models to predict the class label of every pixel of the target CT volume, i.e., the background class and the five mediastinal neoplasm classes. During the training stage, the 2D segmentation models (nnU-Net 2D network, U-Net, and TransUNet) were trained on the 2D axial CT images in the CT volumes, and the nnU-Net 3D full-resolution networks were trained on the whole CT volumes. Dice loss and cross-entropy loss were adopted to train the U-Net and TransUNet models using the AdamW optimizer for 300,000 iterations of batch size 32 at a learning rate of 0.01. For nnU-Net, the training configuration was automatically determined, and the model was trained for 1000 epochs. Since the parameter number of the TransUNet model is large, we trained the models based on the pre-trained model on the ImageNet database[115].

During the inference stage, target CT volumes were first fed into the segmentation models to obtain raw predictions, then the raw predictions were post-processed considering the principle of continuity. Specifically, all the 3D connected components of predicted mediastinal neoplasms were calculated, and those whose z-axis length was less than 3 pixels were considered false positive predictions. Finally, only the largest connected component in each CT volume was predicted as a mediastinal neoplasm. The subtype of the predicted mediastinal neoplasm was determined as the type with the most predicted pixels.

Assessment of similarity between real and synthetic images

We trained autoencoder models to learn representative features of real and synthetic images, and calculated their cosine similarity score. We used 2D U-Net for CT images and 3D U-Net for CT volumes. The models were trained on the internal training set with L1 loss. Parameters were optimized using the AdamW optimizer for 200 epochs at a learning rate of 0.0001. With the U-Net models, we fed CT images into the models and obtained hidden space features from the U-Net encoder network. To reduce the dimension and obtain more precise features, we performed a max-pooling operation for the hidden space features that calculates the maximum value for each channel of the outputs.

Black-box membership inference attack

Generally, well-trained deep learning models have relatively lower losses on the training data than on the test data. This has been successfully utilized to conduct membership inference attacks. In this study, we designed a straightforward attack method based on the model's dice loss on input data. For a target image or volume with ground-truth annotation, we obtained the model's segmentation prediction and calculated the DSC between prediction and ground-truth annotation. If the DSC was greater than a pre-determined threshold, then the attacker inferred that the target image was used to train the model. In this study, we used the optimal threshold of the receiver operator curve as the threshold.

Statistical analysis

To calculate the 95% confidence intervals, we bootstrapped the estimation for 1000 iterations and reported the 2.5th and 97.5th percentiles. We used the two-sided Mann-Whitney-Wilcoxon test to analyze the DSCs of different sizes of mediastinal neoplasms. The analyses are performed using Python packages scikit-learn (version 0.24.2) and statsannotations (version 0.5.0).

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The data that support the findings of this study are divided into two groups: shared data and restricted data. Shared data are available from the manuscript, references, and supplementary materials. Restricted data relating to individuals in this study are subject to a license that allows for the use of the data only for analysis. The internal test datasets, external test datasets, and DiffGuard-generated data will be released upon publication.

Code availability

We have uploaded our code, trained models, and part of the DiffGuard-generated data at https://github.com/ZhanpZhou/DiffGuard (https://doi.org/10.5281/zenodo.13946208). For experiments on nnU-Net, we used the public code at https://github.com/MIC-DKFZ/nnUNet/tree/nnunetv1. For experiments on AsynDGAN, we used the public code at https://github.com/tommy-qichang/AsynDGAN. For DP-SGD, we used the python package pyvacy (version 0.0.23).

Received: 8 May 2024; Accepted: 7 October 2024;
Published online: 20 October 2024

References

1. Esteva, A. et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118 (2017).
2. McKinney, S. M. et al. International evaluation of an AI system for breast cancer screening. *Nature* **577**, 89–94 (2020).
3. Lu, M. Y. et al. AI-based pathology predicts origins for cancers of unknown primary. *Nature* **594**, 106–110 (2021).
4. Song, C., Ristenpart, T. & Shmatikov, V. Machine Learning Models that Remember Too Much. in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security* 587–601 (ACM, Dallas Texas USA, 2017). https://doi.org/10.1145/3133956.3134077.
5. Li, H., Ayache, N. & Delingette, H. Data Stealing Attack on Medical Images: Is It Safe to Export Networks from Data Lakes? *in Distributed, Collaborative, and Federated Learning, and Affordable AI and Healthcare for Resource Diverse Global Health* (eds. Albarqouni, S. et al.) vol. 13573 28–36 (Springer Nature Switzerland, Cham, 2022).
6. Zhu, L., Liu, Z. & Han, S. Deep leakage from gradients. in *Advances in neural information processing systems* (eds. Wallach, H. et al.) vol. 32 (Curran Associates, Inc., 2019).
7. Geiping, J., Bauermeister, H., Dröge, H. & Moeller, M. Inverting Gradients - How easy is it to break privacy in federated learning? in *Advances in neural* information processing systems (eds. Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F. & Lin, H.) vol. 33 16937–16947 (Curran Associates, Inc., 2020).
8. Fredrikson, M., Jha, S. & Ristenpart, T. Model inversion attacks that exploit confidence information and basic countermeasures. in *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security, denver, CO, USA, october 12-16, 2015* (eds. Ray, I., Li, N. & Kruegel, C.) 1322–1333 (ACM, 2015). https://doi.org/10.1145/2810103.2813677.
9. Zhang, Y. et al. The Secret Revealer: Generative Model-Inversion Attacks Against Deep Neural Networks. in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 250–258 (IEEE, Seattle, WA, USA, 2020). https://doi.org/10.1109/CVPR42600.2020.00033.
10. Struppek, L. et al. Plug & play attacks: Towards robust and flexible model inversion attacks. in *International conference on machine learning, ICML 2022, 17-23 july 2022, baltimore, maryland, USA* (eds. Chaudhuri, K. et al.) vol. 162 20522–20545 (PMLR, 2022).
11. Shokri, R., Stronati, M., Song, C. & Shmatikov, V. Membership Inference Attacks Against Machine Learning Models. in *2017 IEEE Symposium on Security and Privacy (SP)* 3–18 (IEEE, San Jose, CA, USA, 2017) https://doi.org/10.1109/SP.2017.41.
12. He, Y., Rahimian, S., Schiele, B. & Fritz, M. Segmentations-Leak: Membership Inference Attacks and Defenses in Semantic Image Segmentation. in *Computer Vision – ECCV 2020* (eds. Vedaldi, A., Bischof, H., Brox, T. & Frahm, J.-M.) vol. 12368 519–535 (Springer International Publishing, Cham, 2020).
13. Zhang, G., Liu, B., Zhu, T., Ding, M. & Zhou, W. Label-Only Membership Inference Attacks and Defenses In Semantic Segmentation Models. *IEEE Trans. Dependable Secure Comput*. 1–1 https://doi.org/10.1109/TDSC.2022.3154029 (2022).
14. Li, N., Qardaji, W., Su, D., Wu, Y. & Yang, W. Membership privacy: a unifying framework for privacy definitions. in *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security* 889–900 (Association for Computing Machinery, New York, NY, USA, 2013). https://doi.org/10.1145/2508859.2516686.
15. Paass, G. Disclosure risk and disclosure avoidance for microdata. *J. Bus. Econ. Stat.* **6**, 487–500 (1988).
16. Shejwalkar, V. & Houmansadr, A. Membership Privacy for Machine Learning Models Through Knowledge Transfer. *Proc. AAAI Conf. Artif. Intell.* **35**, 9549–9557 (2021).
17. Tang, X. et al. Mitigating membership inference attacks by {Self-Distillation} through a novel ensemble architecture. In *31st USENIX security symposium (USENIX security 22)*. 1433–1450 (2022).
18. Nasr, M., Shokri, R. & Houmansadr, A. Machine Learning with Membership Privacy using Adversarial Regularization. in *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security* 634–646 (ACM, Toronto Canada, 2018). https://doi.org/10.1145/3243734.3243855.
19. Hu, H., Salcic, Z., Dobbie, G., Chen, Y. & Zhang, X. EAR: An Enhanced Adversarial Regularization Approach against Membership Inference Attacks. in *2021 International Joint Conference on Neural Networks (IJCNN)* 1–8 (IEEE, Shenzhen, China, 2021). https://doi.org/10.1109/IJCNN52387.2021.9534381.
20. Abadi, M. et al. Deep Learning with Differential Privacy. in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security* 308–318 (ACM, Vienna Austria, 2016). https://doi.org/10.1145/2976749.2978318.
21. Shin, H.-C. et al. Medical Image Synthesis for Data Augmentation and Anonymization Using Generative Adversarial Networks. in *Simulation and Synthesis in* Medical *Imaging* (eds. Gooya, A., Goksel, O., Oguz, I. & Burgos, N.) vol. 11037 1–11 (Springer International Publishing, Cham, 2018).
22. Han, T. et al. Breaking medical data sharing boundaries by using synthesized radiographs. *Sci. Adv.* **6**, eabb7973 (2020).
23. DuMont Schütte, A. et al. Overcoming barriers to data sharing with medical image generation: a comprehensive evaluation. *Npj Digit. Med.* **4**, 141 (2021).
24. Henschke, C. I. et al. CT Screening for Lung Cancer:Prevalence and Incidence of Mediastinal Masses. *Radiology* **239**, 586–590 (2006).
25. Yoon et al. Incidental Anterior Mediastinal Nodular Lesions on Chest CT in Asymptomatic Subjects. *J. Thorac. Oncol. Publ. Int. Assoc. Study Lung Cancer* **13**, 359–366 (2017).
26. Miyazawa, R. et al. Incidental mediastinal masses detected at low-dose CT screening: prevalence and radiological characteristics. *Jpn. J. Radio.* **38**, 1150–1157 (2020).
27. Strollo, D. C., de, C., Melissa, L., Rosado, J. & James, R. Primary Mediastinal Tumors. Part 1*: Tumors of the Anterior Mediastinum. *Chest* **112**, 511–522 (1997).
28. Juanpere, S. et al. A diagnostic approach to the mediastinal masses. *Insights Imaging* **4**, 29–52 (2012).
29. Somepalli, G., Singla, V., Goldblum, M., Geiping, J. & Goldstein, T. Diffusion art or digital forgery? investigating data replication in

diffusion models. in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* 6048–6058 (2023).
30. Yeom, S., Giacomelli, I., Fredrikson, M. & Jha, S. Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting. in *2018 IEEE 31st Computer Security Foundations Symposium (CSF)* 268–282 (IEEE, Oxford, 2018). https://doi.org/10.1109/CSF.2018.00027.
31. Aberle, D. R. et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. *N. Engl. J. Med.* **365**, 395–409 (2011).
32. 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. in *Advances in neural information processing systems 30: Annual conference on neural information processing systems 2017, december 4–9, 2017, long beach, CA, USA* (eds. Guyon, I. et al.) 6626–6637 (2017).
33. Binkowski, M., Sutherland, D. J., Arbel, M. & Gretton, A. Demystifying MMD gans. in *6th international conference on learning representations*, ICLR 2018, vancouver, BC, canada, april 30 - may 3, 2018, conference track proceedings (OpenReview.net, 2018).
34. Salimans, T. et al. Improved techniques for training gans. in *Advances in neural information processing systems 29*: *Annual conference on neural information processing systems 2016, december 5-10, 2016, barcelona, spain* (eds. Lee, D. D., Sugiyama, M., von Luxburg, U., Guyon, I. & Garnett, R.) 2226–2234 (2016).
35. Chang, Q. et al. Synthetic Learning: Learn From Distributed Asynchronized Discriminator GAN Without Sharing Medical Image Data. in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 13853–13863 (IEEE, Seattle, WA, USA, 2020). https://doi.org/10.1109/CVPR42600.2020.01387.
36. Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J. & Maier-Hein, K. H. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* **18**, 203–211 (2021).
37. Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. in *Medical image computing and computer-assisted intervention - MICCAI 2015 - 18th international conference munich, germany, october 5 - 9, 2015, proceedings*, part III (eds. Navab, N., Hornegger, J., I. I. I., W. M. W. & Frangi, A. F.) vol. 9351 234–241 (Springer, 2015).
38. Chen, J. et al. TransUNet: Rethinking the U-Net architecture design for medical image segmentation through the lens of transformers. *Med. Image Anal.* **97**, 103280 (2024).
39. Dwork, C. Differential Privacy. in *Automata, Languages and Programming* (eds. Bugliesi, M., Preneel, B., Sassone, V. & Wegener, I.) vol. 4052 1–12 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2006).
40. Dwork, C. & Roth, A. *The Algorithmic Foundations of Differential Privacy*. (now Publishers Inc, 2013). https://doi.org/10.1561/9781601988195.
41. Gadotti, A., Rocher, L., Houssiau, F., Creţu, A.-M. & De Montjoye, Y.-A. Anonymization: The imperfect science of using data while preserving privacy. *Sci. Adv.* **10**, eadn7053 (2024).
42. Jayaraman, B. & Evans, D. Evaluating differentially private machine learning in practice. in *28th USENIX security symposium, USENIX security 2019, santa clara, CA, USA, august 14-16, 2019* (eds. Heninger, N. & Traynor, P.) 1895–1912 (USENIX Association, 2019).
43. Blanco-Justicia, A., Sánchez, D., Domingo-Ferrer, J. & Muralidhar, K. A Critical Review on the Use (and Misuse) of Differential Privacy in Machine Learning. *ACM Comput. Surv.* **55**, 1–16 (2023).
44. Tayebi Arasteh, S. et al. Preserving fairness and diagnostic accuracy in private large-scale AI models for medical imaging. *Commun. Med.* **4**, 1–12 (2024).
45. Ziller, A. et al. Reconciling privacy and accuracy in AI for medical imaging. *Nat. Mach. Intell*. 1–11 (2024).
46. Choi, E. et al. Generating multi-label discrete patient records using generative adversarial networks. in *Machine learning for healthcare conference* 286–305 (2017).
47. Xie, L., Lin, K., Wang, S., Wang, F. & Zhou, J. Differentially Private Generative Adversarial Network. Preprint at http://arxiv.org/abs/1802.06739 (2018).
48. Baowaly, M. K., Lin, C.-C., Liu, C.-L. & Chen, K.-T. Synthesizing electronic health records using improved generative adversarial networks. *J. Am. Med. Inform. Assoc.* **26**, 228–241 (2019).
49. Zhang, Z., Yan, C., Mesa, D. A., Sun, J. & Malin, B. A. Ensuring electronic medical record simulation through better training, modeling, and evaluation. *J. Am. Med. Inform. Assoc.* **27**, 99–108 (2020).
50. Goodfellow, I. J. et al. Generative adversarial nets. in *Advances in neural information processing systems 27*: *Annual conference on neural information processing systems 2014, december 8-13 2014, montreal, quebec, canada* (eds. Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D. & Weinberger, K. Q.) 2672–2680 (2014).
51. Lin, Y., Wang, Z., Cheng, K.-T. & Chen, H. InsMix: Towards Realistic Generative Data Augmentation for Nuclei Instance Segmentation. in *Medical Image Computing and Computer Assisted Intervention* – MICCAI *2022* (eds. Wang, L., Dou, Q., Fletcher, P. T., Speidel, S. & Li, S.) vol. 13432 140–149 (Springer Nature Switzerland, Cham, 2022).
52. Salehinejad, H., Valaee, S., Dowdell, T., Colak, E. & Barfett, J. Generalization of Deep Neural Networks for Chest Pathology Classification in X-Rays Using Generative Adversarial Networks. in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 990–994 (IEEE, Calgary, AB, 2018). https://doi.org/10.1109/ICASSP.2018.8461430.
53. Ratliff, L. J., Burden, S. A. & Sastry, S. S. Characterization and computation of local Nash equilibria in continuous games. in *2013 51st Annual Allerton Conference on Communication, Control, and Computing (Allerton)* 917–924 (IEEE, Monticello, IL, 2013). https://doi.org/10.1109/Allerton.2013.6736623.
54. Ioffe, S. & Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. in *International conference on machine learning* 448–456 (pmlr, 2015).
55. Goodfellow, I. NIPS 2016 Tutorial: Generative Adversarial Networks. Preprint at http://arxiv.org/abs/1701.00160 (2017).
56. Arora, S., Ge, R., Liang, Y., Ma, T. & Zhang, Y. Generalization and equilibrium in generative adversarial nets (gans). in *International conference on machine learning* 224–232 (PMLR, 2017).
57. Borji, A. Pros and cons of gan evaluation measures. *Comput. Vis. Image Underst.* **179**, 41–65 (2019).
58. Ho, J., Jain, A. & Abbeel, P. Denoising diffusion probabilistic models. in *Advances in neural information processing systems 33*: *Annual conference on neural information processing systems* 2020, NeurIPS 2020, december 6-12, 2020, virtual (eds. Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.-F. & Lin, H.-T.) (2020).
59. Dhariwal, P. & Nichol, A. Q. Diffusion models beat GANs on image synthesis. in *Advances in neural information processing systems 34*: *Annual conference on neural information processing systems 2021, NeurIPS 2021, december 6-14, 2021, virtual* (eds. Ranzato, M., Beygelzimer, A., Dauphin, Y. N., Liang, P. & Vaughan, J. W.) 8780–8794 (2021).
60. Azizi, S., Kornblith, S., Saharia, C., Norouzi, M. & Fleet, D. J. Synthetic data from diffusion models improves ImageNet classification. *Trans. Mach. Learn. Res*.
61. Yang, L., Xu, X., Kang, B., Shi, Y. & Zhao, H. Freemask: Synthetic images with dense annotations make stronger segmentation models. *Adv. Neural Inf. Process. Syst*. **36**, (2024).
62. Tian, Y. et al. Learning vision from models rivals learning vision from data. in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* 15887–15898 (2024).

63. Hammoud, H. A. A. K., Itani, H., Pizzati, F., Bibi, A. & Ghanem, B. SynthCLIP: Are we ready for a fully synthetic CLIP training? In *Synthetic data for computer vision workshop@ CVPR* (2024).

64. Pan, S. et al. 2D medical image synthesis using transformer-based denoising diffusion probabilistic model. *Phys. Med. Biol.* **68**, 105004 (2023).

65. Nguyen, L. X., Sone Aung, P., Le, H. Q., Park, S.-B. & Hong, C. S. A New Chapter for Medical Image Generation: The Stable Diffusion Method. in *2023 International Conference on Information Networking (ICOIN)* 483–486 https://doi.org/10.1109/ICOIN56518.2023.10049010 (2023).

66. Khader, F. et al. Denoising diffusion probabilistic models for 3D medical image generation. *Sci. Rep.* **13**, 7303 (2023).

67. Harb, R., Pock, T. & Müller, H. Diffusion-based generation of histopathological whole slide images at a gigapixel scale. in *Proceedings of the IEEE/CVF winter conference on applications of computer vision* 5131–5140 (2024).

68. Peng, W. et al. Generating Realistic Brain MRIs via a Conditional Diffusion Probabilistic Model. in *Medical Image Computing and Computer Assisted Intervention* – MICCAI *2023* (eds. Greenspan, H. et al.) 14–24 (Springer Nature Switzerland, Cham, 2023). https://doi.org/10.1007/978-3-031-43993-3_2.

69. Dorjsembe, Z., Odonchimed, S. & Xiao, F. Three-dimensional medical image synthesis with denoising diffusion probabilistic models. in *Medical imaging with deep learning* (2022).

70. Xu, X., Kapse, S., Gupta, R. & Prasanna, P. ViT-DAE: Transformer-Driven Diffusion Autoencoder for Histopathology Image Analysis. in *Deep Generative Models* (eds. Mukhopadhyay, A., Oksuz, I., Engelhardt, S., Zhu, D. & Yuan, Y.) 66–76 (Springer Nature Switzerland, Cham, 2024). https://doi.org/10.1007/978-3-031-53767-7_7.

71. Müller-Franzes, G. et al. A multimodal comparison of latent denoising diffusion probabilistic models and generative adversarial networks for medical image synthesis. *Sci. Rep.* **13**, 12098 (2023).

72. Sun, S., Goldgof, G., Butte, A. & Alaa, A. M. Aligning synthetic medical images with clinical knowledge using human feedback. *Adv. Neural Inf. Process. Syst*. **36**, (2024).

73. Takezaki, S. & Uchida, S. An Ordinal Diffusion Model for Generating Medical Images with Different Severity Levels. in *2024 IEEE International Symposium on Biomedical Imaging (ISBI)* 1–5 https://doi.org/10.1109/ISBI56570.2024.10635504 (2024).

74. Ye, J., Ni, H., Jin, P., Huang, S. X. & Xue, Y. Synthetic Augmentation with Large-Scale Unconditional Pre-training. in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023* (eds. Greenspan, H. et al.) 754–764 (Springer Nature Switzerland, Cham, 2023). https://doi.org/10.1007/978-3-031-43895-0_71.

75. Khosravi, B. et al. Synthetically enhanced: unveiling synthetic data's potential in medical imaging research. *eBioMedicine* **104**, 105174 (2024).

76. Reynaud, H. et al. Feature-Conditioned Cascaded Video Diffusion Models for Precise Echocardiogram Synthesis. in *Medical Image Computing and Computer Assisted Intervention* – MICCAI *2023* (eds. Greenspan, H. et al.) 142–152 (Springer Nature Switzerland, Cham, 2023). https://doi.org/10.1007/978-3-031-43999-5_14.

77. Yoon, J. S., Zhang, C., Suk, H.-I., Guo, J. & Li, X. SADM: Sequence-Aware Diffusion Model for Longitudinal Medical Image Generation. in *Information Processing in Medical Imaging* (eds. Frangi, A., de Bruijne, M., Wassermann, D. & Navab, N.) 388–400 (Springer Nature Switzerland, Cham, 2023). https://doi.org/10.1007/978-3-031-34048-2_30.

78. Saeed, S. U. et al. Bi-parametric prostate MR image synthesis using pathology and sequence-conditioned stable diffusion. in *Medical imaging with deep learning* 814–828 (PMLR, 2024).

79. Weber, T., Ingrisch, M., Bischl, B. & Rügamer, D. Cascaded Latent Diffusion Models for High-Resolution Chest X-ray Synthesis. in *Advances in Knowledge Discovery and Data* Mining (eds. Kashima, H., Ide, T. & Peng, W.-C.) 180–191 (Springer Nature Switzerland, Cham, 2023). https://doi.org/10.1007/978-3-031-33380-4_14.

80. Montoya-del-Angel, R., Sam-Millan, K., Vilanova, J. C. & Martí, R. MAM-E: Mammographic Synthetic Image Generation with Diffusion Models. *Sensors* **24**, 2076 (2024).

81. Xu, Y. et al. MedSyn: Text-guided Anatomy-aware Synthesis of High-Fidelity 3D CT Images. *IEEE Trans. Med. Imaging* 1–1 (2024) https://doi.org/10.1109/TMI.2024.3415032.

82. Jiang, L., Mao, Y., Wang, X., Chen, X. & Li, C. CoLa-Diff: Conditional Latent Diffusion Model for Multi-modal MRI Synthesis. in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023* (eds. Greenspan, H. et al.) 398–408 (Springer Nature Switzerland, Cham, 2023). https://doi.org/10.1007/978-3-031-43999-5_38.

83. Zhu, L. et al. Make-A-Volume: Leveraging Latent Diffusion Models for Cross-Modality 3D Brain MRI Synthesis. in *Medical Image Computing and Computer Assisted Intervention* – MICCAI *2023* (eds. Greenspan, H. et al.) 592–601 (Springer Nature Switzerland, Cham, 2023). https://doi.org/10.1007/978-3-031-43999-5_56.

84. Sun, S., Goldgof, G. M., Butte, A. & Alaa, A. M. Aligning Synthetic Medical Images with Clinical Knowledge using Human Feedback.

85. Dorjsembe, Z., Pao, H.-K., Odonchimed, S. & Xiao, F. Conditional Diffusion Models for Semantic 3D Brain MRI Synthesis. *IEEE J. Biomed. Health Inform.* **28**, 4084–4093 (2024).

86. Eschweiler, D. et al. Denoising diffusion probabilistic models for generation of realistic fully-annotated microscopy image datasets. *PLOS Comput. Biol.* **20**, e1011890 (2024).

87. Oh, H.-J. & Jeong, W.-K. DiffMix: Diffusion Model-Based Data Synthesis for Nuclei Segmentation and Classification in Imbalanced Pathology Image Datasets. in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023* (eds. Greenspan, H. et al.) 337–345 (Springer Nature Switzerland, Cham, 2023). https://doi.org/10.1007/978-3-031-43898-1_33.

88. Stojanovski, D., Hermida, U., Lamata, P., Beqiri, A. & Gomez, A. Echo from Noise: Synthetic Ultrasound Image Generation Using Diffusion Models for Real Image Segmentation. in *Simplifying Medical Ultrasound* (eds. Kainz, B. et al.) 34–43 (Springer Nature Switzerland, Cham, 2023). https://doi.org/10.1007/978-3-031-44521-7_4.

89. Zhao, X. & Hou, B. High-fidelity image synthesis from pulmonary nodule lesion maps using semantic diffusion model. in *Medical imaging with deep learning, short paper track*.

90. Xing, X., Papanastasiou, G., Walsh, S. & Yang, G. Less Is More: Unsupervised Mask-Guided Annotated CT image synthesis with minimum manual segmentations. *IEEE Trans. Med. Imaging* **42**, 2566–2576 (2023).

91. Shrivastava, A. & Fletcher, P. T. NASDM: Nuclei-Aware Semantic Histopathology Image Generation Using Diffusion Models. in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023* (eds. Greenspan, H. et al.) 786–796 (Springer Nature Switzerland, Cham, 2023). https://doi.org/10.1007/978-3-031-43987-2_76.

92. Zhuang, Y. et al. Semantic Image Synthesis for Abdominal CT. in *Deep Generative* Models (eds. Mukhopadhyay, A., Oksuz, I., Engelhardt, S., Zhu, D. & Yuan, Y.) 214–224 (Springer Nature Switzerland, Cham, 2024). https://doi.org/10.1007/978-3-031-53767-7_21.

93. Chen, Q. et al. Towards generalizable tumor synthesis. in *Proceedings of the IEEE*/CVF conference on computer vision and pattern recognition 11147–11158 (2024).

94. Huy, P. N. & Minh Quan, T. Denoising Diffusion Medical Models. in *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)* 1–5 https://doi.org/10.1109/ISBI53787.2023.10230674 (2023).

95. Aversa, M. et al. Diffinfinite: Large mask-image synthesis via parallel random patch diffusion in histopathology. *Adv. Neural Inf. Process. Syst*. **36**, (2024).

96. Go, S., Ji, Y., Park, S. J. & Lee, S. Generation of structurally realistic retinal fundus images with diffusion models. in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* 2335–2344 (2024).
97. Macháček, R. et al. Mask-conditioned latent diffusion for generating gastrointestinal polyp images. in *Proceedings of the 4th ACM Workshop on Intelligent Cross-Data Analysis and Retrieval* 1–9 (Association for Computing Machinery, New York, NY, USA, 2023). https://doi.org/10.1145/3592571.3592978.
98. Han, K. et al. MedGen3D: A Deep Generative Framework for Paired 3D Image and Mask Generation. in *Medical Image Computing and Computer Assisted Intervention* – MICCAI *2023* (eds. Greenspan, H. et al.) 759–769 (Springer Nature Switzerland, Cham, 2023). https://doi.org/10.1007/978-3-031-43907-0_72.
99. Thambawita, V. et al. SinGAN-Seg: Synthetic training data generation for medical image segmentation. *PLOS ONE* **17**, e0267976 (2022).
100. Saragih, D. G., Hibi, A. & Tyrrell, P. N. Using diffusion models to generate synthetic labeled data for medical image segmentation. *Int. J. Comput. Assist. Radiol. Surg.* **19**, 1615–1625 (2024).
101. Crespi, L., Loiacono, D. & Sartori, P. Are 3D better than 2D Convolutional Neural Networks for Medical Imaging Semantic Segmentation? in *2022 International Joint Conference on Neural Networks (IJCNN)* 1–8 (IEEE, Padua, Italy, 2022). https://doi.org/10.1109/IJCNN55064.2022.9892850.
102. Wu, J. et al. Medsegdiff-v2: Diffusion-based medical image segmentation with transformer. In *Proceedings of the AAAI conference on artificial intelligence*. vol. 38 6030–6038 (2024).
103. Micikevicius, P. et al. Mixed precision training. in *6th international conference on learning representations*, ICLR 2018, Vancouver, BC, Canada, april 30–may 3, 2018, conference track proceedings (OpenReview.net, 2018).
104. Song, J., Meng, C. & Ermon, S. Denoising diffusion implicit models. in *9th international conference on learning representations*, ICLR 2021, virtual event, austria, may 3–7, 2021 (OpenReview.net, 2021).
105. Bao, F., Li, C., Zhu, J. & Zhang, B. Analytic-DPM: an analytic estimate of the optimal reverse variance in diffusion probabilistic models. in *The tenth international conference on learning representations, ICLR 2022, virtual event, april 25-29, 2022* (OpenReview.net, 2022).
106. Rombach, R., Blattmann, A., Lorenz, D., Esser, P. & Ommer, B. High-Resolution Image Synthesis with Latent Diffusion Models. in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 10674–10685 (IEEE, New Orleans, LA, USA, 2022). https://doi.org/10.1109/CVPR52688.2022.01042.
107. Arora, A. & Arora, A. Synthetic patient data in health care: a widening legal loophole. *Lancet* **399**, 1601–1602 (2022).
108. Appenzeller, A., Leitner, M., Philipp, P., Krempel, E. & Beyerer, J. Privacy and utility of private synthetic data for medical data analyses. *Appl. Sci.* **12**, 12320 (2022).
109. Giuffrè, M. & Shung, D. L. Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. *Npj Digit. Med.* **6**, 1–8 (2023).
110. Teo, C., Abdollahzadeh, M. & Cheung, N.-M. M. On measuring fairness in generative models. *Adv. Neural Inf. Process. Syst*. **36**, (2024).
111. Chen, R. J., Lu, M. Y., Chen, T. Y., Williamson, D. F. K. & Mahmood, F. Synthetic data in machine learning for medicine and healthcare. *Nat. Biomed. Eng.* **5**, 493–497 (2021).
112. Roden, A. C. et al. Distribution of mediastinal lesions across multi-institutional, international, radiology databases. *J. Thorac. Oncol.* **15**, 568–579 (2020).
113. Marx, A. et al. The 2015 World Health Organization Classification of Tumors of the Thymus: Continuity and Changes. *J. Thorac. Oncol.* **10**, 1383–1395 (2015).
114. Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. in *7th international conference on learning representations*, ICLR 2019, new orleans, LA, USA, may 6-9, 2019 (OpenReview.net, 2019).
115. Deng, J. et al. ImageNet: A large-scale hierarchical image database. in *2009 IEEE Conference on Computer Vision and Pattern Recognition* 248–255 (IEEE, Miami, FL, 2009). https://doi.org/10.1109/CVPR.2009.5206848.

Acknowledgements

This study was supported by the National Key R&D Program of China (2018YFA0704000), the National Natural Science Foundation of China (61822111, 62021002), and the Zhejiang Provincial Natural Science Foundation (LDT23F02024F02). This study was also supported by Institute for Brain and Cognitive Science, Tsinghua University (THUIBCS) and Beijing Laboratory of Brain and Cognitive Intelligence, Beijing Municipal Education Commission (BLBCI).

Author contributions

Z.Z., Y.G., and F.X. contributed to the conceptualization. H.L. and J.H. supervised the data collection. Z.Z. and R.T. processed the collected data. Z.Z. developed the algorithm, conducted the experiments and performed the analysis. Z.Z. and Y.G. wrote the manuscript. Y.G. and F.X. supervised the study. All authors have read and approved the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-024-01290-7.

Correspondence and requests for materials should be addressed to Yuchen Guo or Feng Xu.