Repeat and Concatenate: 2D to 3D Image Translation with 3D to 3D Generative Modeling

Abril Corona-Figueroa, Hubert P. H. Shum, Chris G. Willcocks
Department of Computer Science, Durham University, Durham, UK
https://github.com/abrilcf/3D-3D_repeat-concatenate

Abstract

This paper investigates a 2D to 3D image translation method with a straightforward technique, enabling correlated 2D X-ray to 3D CT-like reconstruction. We observe that existing approaches, which integrate information across multiple 2D views in the latent space, lose valuable signal information during latent encoding. Instead, we simply repeat and concatenate the 2D views into higher-channel 3D volumes and approach the 3D reconstruction challenge as a straightforward 3D to 3D generative modeling problem, sidestepping several complex modeling issues. This method enables the reconstructed 3D volume to retain valuable information from the 2D inputs, which are passed between channel states in a Swin UNETR backbone. Our approach applies neural optimal transport, which is fast and stable to train, effectively integrating signal information across multiple views without the requirement for precise alignment; it produces non-collapsed reconstructions that are highly faithful to the 2D views, even after limited training. We demonstrate correlated results, both qualitatively and quantitatively, having trained our model on a single dataset and evaluated its generalization ability across six datasets, including out-of-distribution samples.

1. Introduction

2D to 3D image translation is a class of computer vision problems where the goal is to learn the mapping between one or more 2D images and a corresponding 3D volumetric image. Years of research in this topic have given rise to many image and graphics applications, such as augmented reality [7, 27], sensor fusion [25, 33], scene rendering [5, 19], and multimodal translation [17, 38, 50]. The latter has attracted attention in the medical domain, including 2D X-ray, Computed Tomography (CT), Magnetic Resonance Imaging (MRI), Ultrasound, among others, due to the potential to translate from cheap, low-quality and available medical imaging modalities to those that are more expensive, with high-waiting times, or exhibit harmful ionizing radiation. While translating between image modalities of common dimensionality (i.e., 2D to 2D, 3D to 3D) has achieved notable results [50] due to its seemingly straightforward adaptation of specific state-of-the-art (SOTA) deep learning models, translating between data of distinct dimensionality poses further challenges. In particular, converting an image from lower-dimension to a higher-dimensional representation (e.g., 2D to 3D) is considered a reconstructive generative modelling problem [4, 45].

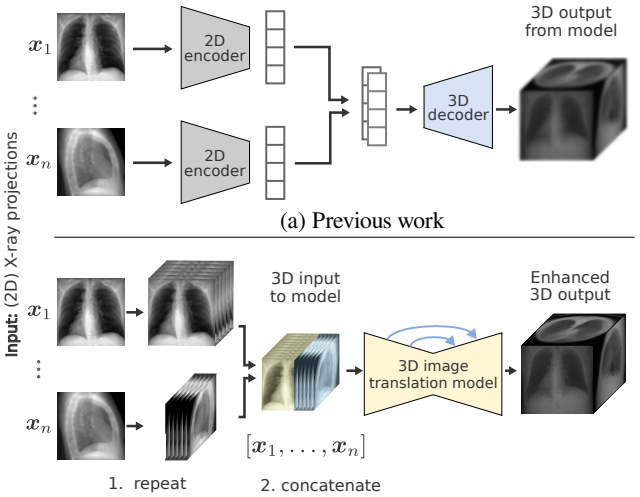


Figure 1. (a) Previous approaches focus on 2D to 3D mapping, often employing asymmetric architectures and compressed latent encoding. (b) In contrast, we propose 3D to 3D mapping from repeated and concatenated inputs, enabling faster training with highly correlated outputs without latent compression, even with small datasets (a few hundred images).

One highly relevant application in the medical field involves obtaining 3D CT representations from 2D X-ray projections [30]. However, unlike image super-resolution approaches, bridging the 2D to 3D data dimensionality gap presents a unique modeling challenge to estimate spatial missing details [45]. Moreover, machine learning models

achieve remarkable results thanks to abundant natural image data, but medical models often struggle due to the limited size of medical datasets. Furthermore, real-world medical 3D datasets involve volumetric features with varying density structures, making hallucinated data a significant concern that can prove counterproductive [34].

Existing methods approach 2D to 3D medical image translation through asymmetrical architectures and/or incorporate various regularization techniques to enforce plausible reconstructions [35, 49] (Fig. 1). While these techniques may generate high-quality outputs, even with high-frequency details, they often lack correlation to the input data. This means that the model becomes overly reliant on the latent prior, potentially ignoring the original 2D signal. The situation is further exacerbated if the training data is limited (few hundred images), so when the model is presented with inputs significantly different from the training dataset, such as out-of-distribution samples, it is likely to perform poorly making it unusable for practical scenarios.

In this paper, we propose a simple 2D to 3D mapping translation framework that addresses the correlation issue, ensuring highly-associated outputs even in limited datasets. We achieve this by a preprocessing step that preserves 2D information content throughout the network transformations without relying on other priors. First, we repeat the various 2D input projections to match the output 3D target depth (Fig. 1). These 3D volumes are then concatenated into a single higher-channel 3D volume. Then we propose a 3D to 3D conditional generative modeling approach that applies neural optimal transport with the de-biased Sinkhorn divergence. Such an approach effectively allows mass splitting; we find this particularly suitable for our task, as the original 2D inputs contain valuable information content, requiring each area of the 2D inputs to contribute to entire 3D regions of details in the synthesized 3D image (Fig. 2).

The approach was found to give significant improvement in terms of generalization, while being stable to train with only a few hundred images (ideal for medical/real-world datasets). Our evaluation showed that the method generates plausible reconstructions after only 2,000 training optimization steps, which can generalize to out-of-distribution inputs after approximately 28 hours of training.

To summarize, our contributions are

1. An alternative and simple 2D to 3D image translation framework that effectively reconstructs a CT volumetric representation given only one or any number of X-ray projections, with potential application in clinical use (e.g., cutting down costs and radiation to patients).
2. A processing pipeline that retains all 2D information throughout the 3D transformation network without loosing information content in the latent encoding.
3. Demonstrable generalization with very limited data (a few hundred images) trained in less than 28 hours. The processing step and approach can be directly applied to other generative modelling approaches and applications.

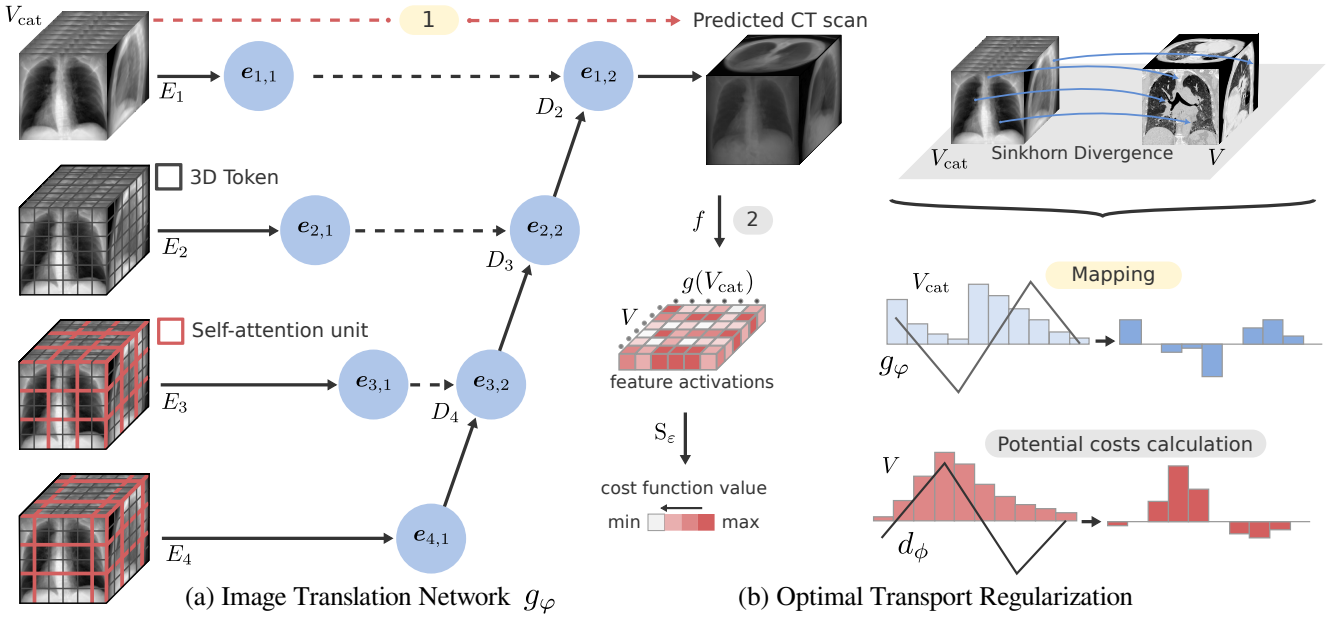
2. Related Work

2D to 3D image translation: Several methods have attempted to obtain 3D representations from 2D images for medical applications, such as CT reconstruction from X-rays. Conditional approaches, generating 3D images from one or two 2D images are often based on Generative Adversarial Networks (GANs) [13] and Variational Autoencoders (VAEs) [22]. These methods transform noise sampled from a Gaussian prior, which is concatenated with a latent representation of the 2D data; this may loose or hallucinate important 2D signal information not captured in the latent encoding. Strategies to mitigate this include incorporating 3D priors from real data [18] or simply generating 2D slices as output, which are stacked into a 3D representation [52].

Another line of work involves asymmetric autoencoder architectures, where a 2D encoder extracts spatial features, which are then expanded to fit into a 3D decoder [15, 24, 26, 32, 40]. To mitigate information loss at the bottleneck, additional objectives are learned, enforcing a more interpretable latent space or adding an additional 2D encoder [49]. Recent approaches explore implicit neural representations [5, 28, 39] or two-stage vector-quantized approaches combined with diffusion [6]. In contrast, we explore a simpler approach that generalizes from small datasets (~1k images) in a 3D to 3D setting without relying on more complex asymmetric architectures.

Regularization and prior knowledge: Generalizing beyond the examples in the training set is a fundamental aspect of machine learning models for medical applications. Key problems are mode collapse and overfitting, which arise due to the limited availability and real-world complexity of medical datasets. This is further exacerbated by data noise and heterogeneity among different imaging modalities. Fortunately, regularization techniques such as parameter constraints or augmentation can mitigate overfitting [31, 41] and help alleviate data shortages.

Data augmentation, in particular, is an established solution to mitigate overfitting [11, 21] by introducing semantic-preserving data transformations such as rotation, translation, and contrast adjustments. However, selecting and combining augmentations is a delicate process, especially to prevent undesirable outcomes such as the model learning to generate the augmented distribution and deviating from its original objective [20]. In this context, incorporating prior knowledge from initial assumptions about the data distribution can guide learning for improved generalization, such as additional conditional information in generative modeling or via additional terms to the main objective functions [10, 44, 46]. In our approach, we know that 2D X-rays cap-



(a) Image Translation Network g_φ (b) Optimal Transport Regularization

Figure 2. Proposed 2D to 3D image translation approach. (a) We learn the mapping between 2D inputs and their corresponding 3D representation adapting a Swin UNe-Transformer g_φ [16] for image translation. (b) Optimization is based on the dual optimal transport regularization between networks g_φ and d_ϕ, comparing data points from the two image spaces using the Sinkhorn divergence S_ε on the activations of a feature extractor f.

ture the underlying 3D objects over a range of depths, acting as a prior that motivates our repeat and concatenate mapping strategy. We find this facilitates improved generalization for both in- and out-of-distribution inputs, even when trained on small datasets.

Deep generative modeling: 2D to 3D image translation involves modeling the probability of unobserved 3D regions which is therefore a generative modeling problem. Deep generative modeling (DGM) is a very large field, from which mainstream approaches include GANs, VAEs, normalizing flows, autoregressive models and probablistic diffusion models [4]. These methods classically balance an empirically observed trilemma of modeling quality, mode coverage and the number of 'steps' taken during sampling [47]. VAEs and normalizing flows are well-known for lower quality sampling, while GANs suffer from mode collapse, capturing only part of the distribution during training. Various strategies, such as weight clipping, gradient penalties, and spectral normalization, have been proposed to mitigate these issues, but they often fail to address fundamental challenges rooted in optimal transport (OT) theory.

Neural optimal transport: In contrast, neural optimal transport offers a more nuanced approach for mapping probability distributions, which is especially relevant for handling the intricacies of medical imaging data. Traditional methods like f-divergences, commonly employed in GANs, perform poorly with smaller, high-dimensional datasets, failing to provide meaningful gradients for effective training [42]. Capturing variability in limited medical data is crucial, and leveraging regularity in human datasets is one approach. For example, organs and bones consistently occupy locations in the human chest, allowing neural networks to learn the mapping between X-ray and 3D CT volumes. OT is relevant in this setting by providing a rigorous distance metric that respects the geometric properties of the distributions. It computes the minimal cost required to transform one distribution into another, offering a coherent and robust measure of similarity that is both intuitive and stable [8]. This geometrically-aware approach enables a more direct and meaningful comparison of medical images across different modalities, potentially facilitating the development of more accurate and reliable diagnostic tools.

3. Method

We initially detail our 2D to 3D processing approach (Fig. 1), and then discuss our 3D to 3D generative modeling regularized with optimal transport (Fig. 2). We also justify how the 3D to 3D mapping approach improves correlation by analyzing the information content through the transformation strategies.

3.1. Processing pipeline

At a high-level, we wish to investigate whether multi-view 2D information can be combined for accurate 3D synthesis without latent encoding. Previous approaches integrate multi-view information in the latent space, as it typically captures a degree of spatial and rotational invariance. However the latents, even with information-rich modeling, fail to retain the full signal from the input 2D views. While, through modern deep generative modeling, they may synthesize high-quality signals with high-frequency details, these are typically prone to over-hallucination [6] which can potentially pose fatal in real-world medical application. In this application, we would rather have blurriness and uncertainty at the expense of outputs that are highly-correlated to their corresponding 2D inputs.

To achieve this objective, we propose increasing the dimensionality at the 2D inputs to ensure the signal can influence an entire region of the 3D volume, and concatenating these inputs ensuring that information content loss is reduced within 3D to 3D residual architectures.

More formally, we outline our processing pipeline for N input X-ray projections. Let $I \in \mathbb{R}^{C\times H\times W}$ be an input 2D X-ray, representing a single view with height H, width W and channels C. For a set of N views, we have $\{I^i\}_{i=1}^N$, each aiming to contribute to the reconstruction of a 3D CT volume $V \in \mathbb{R}^{C\times H\times W\times D}$, where D is the depth.

We 'stretch' the 2D inputs to match the dimensions of the target 3D volume; we repeat each view D times by $\Gamma : \mathbb{R}^{C\times H\times W} \rightarrow \mathbb{R}^{C\times H\times W\times D}$ and coarsely align them by transposing views that differ by 90-degrees (leaving the others unchanged, see our experiments section on *Views alignment* for further discussion on this). This is applied for each of the N views, which are concatenated across the depth axis (denoted by $\oplus$) yielding a composite volume $V_{\text{cat}} \in \mathbb{R}^{NC\times H\times W\times D}$, where

$$V_{\text{cat}} = \bigoplus_{i=1}^{N} \Gamma(I^i), \tag{1}$$

and the reconstructed 3D volume is modeled by a U-Net based architecture $g_\varphi^{\text{unet}} : \mathbb{R}^{NC\times H\times W\times D} \rightarrow \mathbb{R}^{C\times H\times W\times D}$.

3.2. Mapping network and generative modeling

For our 3D to 3D mapping network g_φ with parameters φ we considered a residual U-Net [36] equipped with self-attention following the Swin UNEt-TRansformer (Swin UNETR) model [16], applied for image translation instead of image segmentation. Based on experimenting with a variety of off-the-shelf U-Net backbones, we settled on Swin UNETR as it empirically generated higher image quality outputs across all metrics.

3.2.1 Modeling with neural optimal transport

For the generative modeling task, we applied the dual regularized optimal transport (OT) learning approach [14] using the geometric de-biased Sinkhorn divergence $\mathrm{S}_\varepsilon(\cdot)$ [9] as the cost function. The Sinkhorn divergence is suitable for this task as it is an efficient yet positive and definite approximation of OT that interpolates between Maximum Mean Discrepancies (MMD) and OT. Specifically, we incorporate a feature extractor f that reduces the dimensionality of $g(V_{\text{cat}})$ and V through a mapping into a reduced dimensional vector as suggested in [12]. We parametrized f as a neural network and calculate the cost function in the latent space,

$$\mathrm{S}_\varepsilon(\alpha,\beta) := \mathrm{OT}_\varepsilon(\alpha,\beta) - \frac{1}{2}\mathrm{OT}_\varepsilon(\alpha,\alpha) - \frac{1}{2}\mathrm{OT}_\varepsilon(\beta,\beta), \tag{2}$$

where α represents the distribution of predicted features and β the distribution of ground truth 3D CT features extracted from f. While calculating the cost in the data space might suffice in some cases, we find that incorporating the feature extractor f is beneficial with 3D data as it reduces its dimensionality and makes our mapping network less susceptible to mode collapse.

This modeling approach is similar to the min-max objective in GANs where the Swin UNETR mapping network g_φ (Fig. 2a) replaces the generator and a residual discriminator d_ϕ assigns transportation costs to the produced samples $g_\varphi(V_{\text{cat}})$. We trained our networks using:

$$\mathcal{L}_g = \mathbb{E}_{(V_{\text{cat}},V)\sim p_{\text{data}}}\left[\mathrm{S}_\varepsilon(g(V_{\text{cat}}),V) - \lambda\, d(g(V_{\text{cat}}))\right], \tag{3}$$

$$\mathcal{L}_d = \mathbb{E}_{(V_{\text{cat}},V)\sim p_{\text{data}}}\left[d(g(V_{\text{cat}})) - d(V)\right]. \tag{4}$$

The optimization of the network d involves finding a function that yields the minimal cost associated with transporting mass from each point in V_{cat} to each point in V. Overall this allows us to align the distributions of features from ground truth CT samples with our samples from g_φ attained from the concatenated volumes from X-ray views V_{cat}, through regularized dual OT (Fig. 2b).

Even though we can use the L2 distance to approximate the OT plan, it does not capture the underlying structure of the distributions or how to transport mass from one to another. In contrast, the Sinkhorn divergence induces a scaling process that ensures, at each step, the resulting transportation plan satisfies probability distribution constraints, thus helps avoiding common training instability issues in adversarial training. Furthermore, our model relies solely on 2D views, promoting deterministic mapping and reducing hallucination by avoiding sampling from additional distributions, such as Gaussian, typical of multimodal image translation. We find that OT regularization greatly reduces overfitting but may introduce blur in the generated outputs due to the unconstrained nature of 2D-3D translation.

3.3. Justification

We justify our processing pipeline by examining information loss through different approaches. In previous works, encoded features $\mathbf{z} \in \mathbb{R}^M$ are typically obtained through an encoding function $\mathbf{z} = f^{\text{enc}}(I)$, where typically, $M < CHW$, indicating a reduction in dimensionality. The decoding function $V = f^{\text{dec}}(\mathbf{z})$ similarly tries to reconstruct a volume given the latent.

Information content $\mathrm{I}(\cdot)$ Let $\mathrm{I}(\cdot)$ quantify the informational content of an image or volume, indicating that $\mathrm{I}(V)$ is maximized when V contains the full detail and structure inherent to the original 3D object.

We show that the repetition and concatenation approach retains information when integrated via U-Net over multiple views, compared to classicial encoder-decoder approaches

$$g^{\text{unet}}(V_{\text{cat}}) \succ f^{\text{dec}}\left(\oplus_{i=1}^{N} f^{\text{enc}}(I^i)\right),$$

where $\succ$ denotes higher fidelity (closeness to original) in 3D reconstruction, directly correlated with informational content $\mathrm{I}(\cdot)$.

Information loss in encoding and decoding Dimensionality reduction through encoding implies: $M < CHW$, highlighting a reduction in the capacity to represent the full informational content of the original volume. This reduction leads to inherent information loss, which the decoding process cannot fully recover as

$$M < CHW \implies \mathrm{I}(f^{\text{dec}}(\mathbf{z})) < \mathrm{I}(I), \tag{5}$$

due to the lossy nature of compression in f^{enc} and the limited ability of f^{dec} to fully recover original information.

Information retention through skip-connections The skip connections in g^{unet} enable direct transfer of features across layers, preserving and refining informational content where

$$\mathrm{I}(g^{\text{unet}}(V_{\text{cat}})) \approx \mathrm{I}(V_{\text{cat}}) = \mathrm{I}(I), \tag{6}$$

especially when V is constructed from repeating I across the depth dimension (without information loss), thereby maintaining high fidelity from input to output.

Information retention across views The concatenation of repeated views tends to preserve informational content, while the concatenation of encoded views may lead to information loss due to aggregation

$$\mathrm{I}\left(\oplus_{i=1}^{N} \Gamma(\boldsymbol{x}^i)\right) \approx \sum_{i=1}^{N} \mathrm{I}(\Gamma(I^i)), \tag{7}$$

LIDC-IDRI dataset [1] (in-of-distribution inputs) using two views

Experiment	↑ SSIM	↑ PSNR	↓ MSE	↓ MAE
2D-3D AE	0.1545	3.804	995.2582	5.81
2D-3D AE + L2 Norm.	0.2086	10.833	197.2649	2.5103
3D-3D U-Net	**0.4787**	**22.063**	**58.7546**	**0.9264**

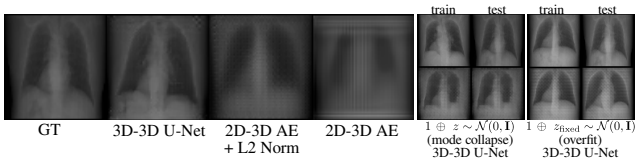


Figure 3. Effect of reformulating 2D-3D mapping into 3D-3D.

$$\mathrm{I}\left(\oplus_{i=1}^{N} g^{\text{enc}}(\boldsymbol{x}^i)\right) \leq \sum_{i=1}^{N} \mathrm{I}(I^i). \tag{8}$$

Therefore, the combination of multiple views through concatenation and subsequent processing with g^{unet} tends to retain information content, whereas concatenating the encodings leads to significant loss in the correlation on aggregate, leading to

$$\mathrm{I}\left(g^{\text{unet}}\left(\oplus_{i=1}^{N} \Gamma(I^i)\right)\right) > \mathrm{I}\left(f^{\text{dec}}\left(\oplus_{i=1}^{N} f^{\text{enc}}(I^i)\right)\right), \tag{9}$$

indicating the potential of enhanced fidelity of 3D reconstruction achieved through a U-Net based architecture with repeated concatenation and skip connections over latent integration frameworks.

4. Experiments

We trained our models on the LIDC chest dataset [1], which consists of only 916 training CT scans, and tested them on both in-distribution and out-of-distribution inputs from six lung datasets. We created paired datasets generating digitally reconstructed radiographs as 2D inputs from the CT scans using the open-source software Plastimatch, following previous work [5, 6, 35, 49]. Our model's convergence plateaued after 2,000 iterations, taking only an average of 28 hours to reach 5,000 steps (selected weight). The training runs were performed on NVIDIA TITAN RTX with a batch size of eight using PyTorch.

Mapping reformulation In Figure 3, we present results from initial experiments comparing an asymmetrical 2D to 3D strategy versus our reformulation as a 3D to 3D mapping. The 2D to 3D approach involves aggregating the individual encoded 2D view features, which serve as input to a 3D decoder. However, this results in highly blurred outputs due to the information content loss from the latent encoding. Despite the simplicitiy of such and approach in terms of feature alignment, important fine-grained details are missing. In contrast, our approach reformulates the problem into a 3D to 3D mapping in a simple architecture, achieving high-fidelity image translations even when using a single input

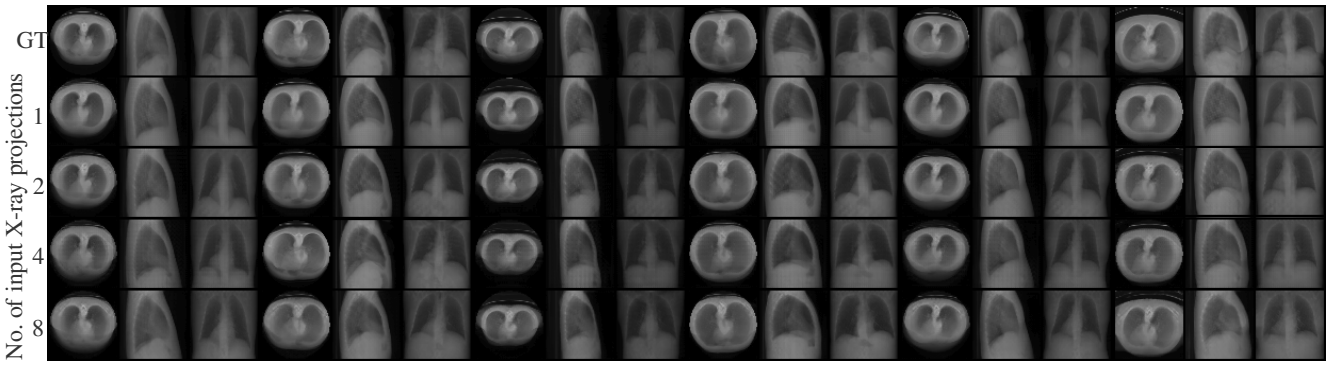


Figure 4. Example projections from generated 3D CT volumes from inputs using one, two, four and eight X-rays; obtained from our 3D-3D translation approach with Swin UNETR backbone. We use testing instances from LIDC-IDRI dataset [1].

In-distribution inputs (LIDC-IDRI dataset [1])

No. input views	↑ SSIM	↑ PSNR	↓ MSE	↓ MAE
$1 \oplus z \sim \mathcal{N}(0, \mathbf{I})$	0.2337	20.2827	0.0403	0.1342
1	0.4891	23.2198	0.0214	0.0751
2	0.5272	24.3192	0.0180	0.0665
4	0.5129	24.3358	0.0167	0.0666
8	**0.5402**	**24.6352**	**0.0155**	**0.0613**

Out-of-distribution inputs (MIDRC-1b dataset [43])

$1 \oplus z \sim \mathcal{N}(0, \mathbf{I})$	0.1353	15.3223	0.0800	0.2334
1	0.4043	20.0800	0.0526	0.1510
2	**0.4048**	20.3636	0.0505	0.1474
4	0.3779	**21.2115**	**0.0439**	**0.1337**
8	0.2029	17.7972	0.0777	0.2029

Table 1. Primary results of our mapping approach with varying input views. Each model was trained on LIDC dataset [1] and tested for both in- and out-of-distribution inputs from MIDRC-1b dataset [43]. We computed metrics five times with different random seeds and report their average.

view. We ablate our 'repeat and concatenate' preprocessing pipeline by instead concatenating noise sampled from a normal distribution. When the vector $\boldsymbol{z}$ is fixed for each patient, the model causes overfitting, while gathering a new $\boldsymbol{z}$ for each iteration results in mode collapsing, where it produces the same output irrespective of the input.

Ablations & comparisons Our approach allows the use of multiple views as input without modifying our model's architecture. We observe that the model benefits from additional views when tested on in-distribution inputs (Fig. 4). However, this correlation was found not strictly hold for out-of-distribution inputs. This suggests that when there are disparities in view alignment or style-domain features compared to the training distribution, the model predominantly relies on views that closely resemble the learned patterns. Investigating whether certain projections contain more valuable information than others is a valuable research direction for future work. Overall, we find that the combination of 2 or 4 views to be empirically effective (Table 1).

(a) In-of-distribution inputs

Dataset	**Method**	↑ SSIM	↑ PSNR	↓ MSE	↓ MAE
LIDC-IDRI [1]	X2CT-GAN [49]	0.321	19.68	0.045	0.151
	CCX-rayNet [35]	0.386	22.66	0.032	0.108
	Ours	**0.527**	**24.35**	**0.018**	**0.066**

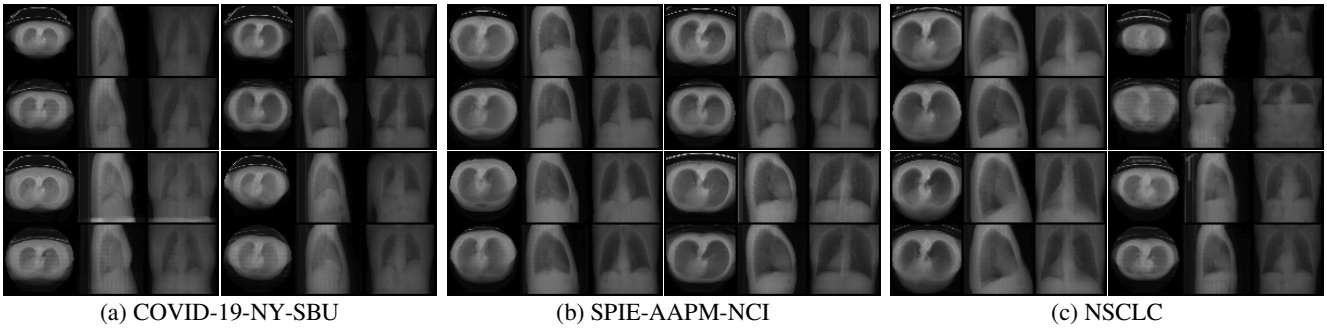
(b) Out-of-distribution inputs

COVID-19-NY-SBU [37]	X2CT-GAN [49]	0.236	16.74	0.089	0.199
	CCX-rayNet [35]	0.205	19.03	0.054	0.144
	Ours	**0.400**	**22.78**	**0.022**	**0.077**
SPIE-AAPM-NCI [2]	X2CT-GAN [49]	0.174	15.63	0.115	0.245
	CCX-rayNet [35]	0.130	17.87	0.080	0.196
	Ours	**0.399**	**22.04**	**0.026**	**0.087**
MIDRC-1b [43]	X2CT-GAN [49]	0.220	14.70	0.156	0.360
	CCX-rayNet [35]	0.114	18.06	0.076	0.228
	Ours	**0.404**	**20.36**	**0.050**	**0.147**
ANTI-PD-1 [29]	X2CT-GAN [49]	0.286	18.04	0.072	0.164
	CCX-rayNet [35]	0.205	18.15	0.067	0.167
	Ours	**0.349**	**20.93**	**0.040**	**0.112**
LCTSC [48]	X2CT-GAN [49]	0.326	19.20	0.052	0.125
	CCX-rayNet [35]	0.206	19.75	0.062	0.156
	Ours	**0.331**	**20.38**	**0.040**	**0.109**
NSCLC [3]	X2CT-GAN [49]	0.280	18.13	0.072	0.163
	CCX-rayNet [35]	0.122	16.62	0.093	0.216
	Ours	**0.315**	**19.85**	**0.047**	**0.126**

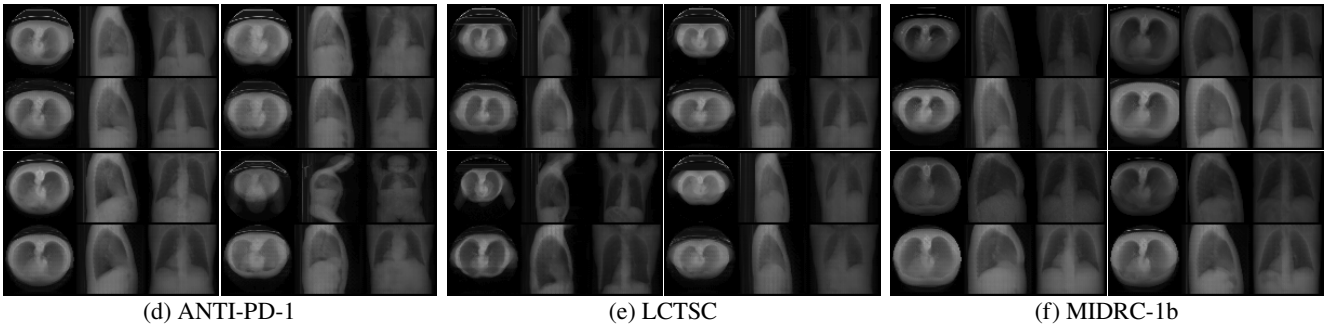
Table 2. Quantitative results on both in-distribution and several out-of-distribution datasets using only two input X-rays. We used our model weights for 5,000 training optimization steps, trained with Swin-U-Net. Other models weights are from 100 epochs (~ 90k iterations).

In Table 2 we compare our approach to paired alternative methods in terms of quality of the 3D outputs for both in- and out-of-distribution inputs. Despite variations in datasets originating from different imaging systems, resolutions, and patients' health conditions, our model demonstrates superior performance across all metrics and maintains consistency across all out-of-distribution datasets. Refer to Figures 5 and 9 for qualitative results.

Gradients & patient-specific learning In Figure 8, we visualize the magnitudes of the gradients of our model with respect to a varying number of input views. We observe higher variability across different patients, indicated by the color variations for one, two, four, and eight views independently. On the other hand, if we analyze intra-patient



(a) COVID-19-NY-SBU (b) SPIE-AAPM-NCI (c) NSCLC



(d) ANTI-PD-1 (e) LCTSC (f) MIDRC-1b

Figure 5. CT projections from generated 3D volumes on various out-of-distribution lung datasets. Model weights were selected from iteration 5,000 on the LIDC-IDRI dataset [1]. GT projections are displayed in odd rows, while our model's outputs are shown in even rows.

(a) In-of-distrib. inputs (LIDC-IDRI dataset [1])

Oblique views ablation	↑ SSIM	↑ PSNR	↓ MSE	↓ MAE
no alignment	**0.5001**	**24.0513**	**0.0178**	**0.0682**
coarse alignment	0.4248	22.5182	0.0259	0.0872

(b) Out-of-distrib. inputs (COVID-19-NY-SBU dataset [37])

no alignment	**0.3581**	**21.9844**	**0.0272**	**0.0881**
coarse alignment	0.3365	21.0851	0.0340	0.1026

Figure 6. Results on oblique views with coarse alignment.

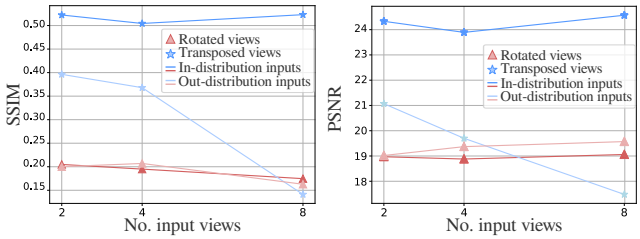


Figure 7. Comparison between only transposing perpendicular views (indicated by a star marker) and a coarse alignment (triangle marker), using two, four, and eight input views. Smoother colors represent results for out-of-distrib. inputs (MIDRC dataset [43]).

gradients, we notice lower variability, suggesting consistent patterns captured in views coming from the same patient. With our approach, the generative model implicitly learns patient-specific representations, where outputs reflect differences in anatomical structures and/or pathological conditions present in the input data from each patient.

View alignment We study the effect of aligning the input views during the preprocessing stage prior to the repeat and concatenation operations. Initially, we experimented without any alignment, which resulted in outputs containing checkerboard-like artifacts that diminished after an additional 1k training iterations. However, a simple transposition of perpendicular views from coronal and sagittal planes alleviated this. To test our model capacity on views outside of such geometry, we tested on four oblique projections, for which we find that our model performs equally well without relying on any sort of alignment (Fig. 6). We also investigated a coarse alignment by approximating their locations through 3D rotations using data transformations from Kornia's library (Figs. 6, 7). However, we observe that this might required a precise and potentially non-affine transformation which is non-trivial. Overall, we find that the model learns some invariance to this, where a simple transpose consistently yields improved results without apparent artifacts (Fig. 7). While precise projection alignment in 3D space could potentially enhance results, we consider there may be a tradeoff between alignment robustness (e.g., due to patient movement) and generation accuracy. In the future, an iterative 3D alignment approach would be worthwhile investigating as a secondary objective.

4.1. Implementation Details

Our training algorithm is based on the neural optimal transport approach by Korotin *et al.* [23]. In this approach,

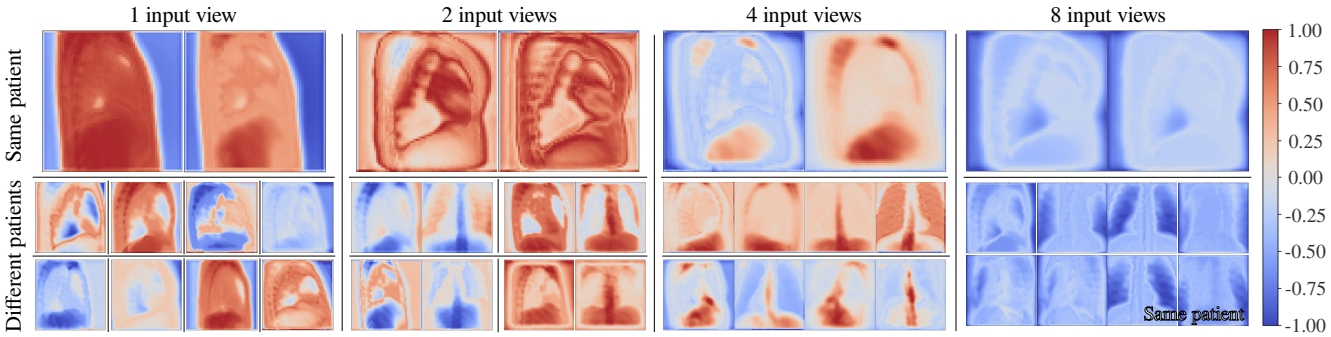


Figure 8. Gradient magnitudes with respect to the input views (X-rays) of the proposed image translation model. The *upper row* displays the variability within slices extracted from the 3D input volume of the same patient, while the *bottom row* shows the attribution among slices from different patients.

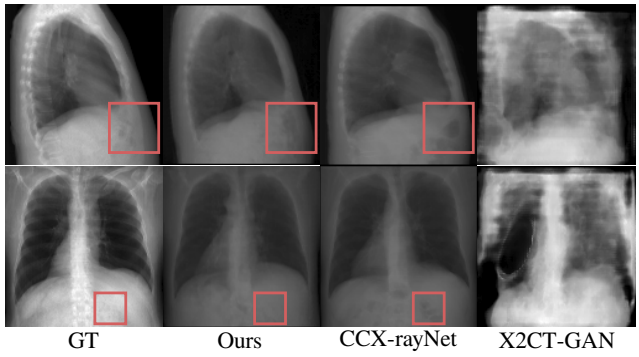


Figure 9. Examples of the correlated 3D projections from our proposed approach compared to alternative supervised methods X2CT-GAN [49] and CCX-rayNet [35].

the mapping network g_φ undergoes $k = 10$ iterations while the parameters of the potential network d_ϕ remain frozen. Then, a single training step is performed for d, unfreezing its parameters while freezing those of g. Our feature extractor f is parameterized by a 3D convolutional neural network with Kaiming weight initialization. We set $\lambda = 0.1$ to control the degree of regularization of the reconstructed outputs by d. Improving generation quality could be achieved by scaling to higher resolutions and adjusting the λ parameter. Our networks are trained using the AdamW optimizer with parameters $\beta_1 = 0.5$, $\beta_2 = 0.999$, $\epsilon = 0.001$, and a learning rate of 10^{-5} with weight decay. We apply a combination of differential augmentations [51], including random contrast, rotation, and horizontal flip, to the input 2D X-rays in all experiments. 3D reconstructions are saved in both numpy and .mha formats for visualization using medical image software such as 3D Slicer.

5. Limitations and future directions

The approach, while simple and intuitive, has several clear limitations. Firstly, our current architecture attempts the non-linear transformation in a 'single' transformative step, resulting in uncertainty and therefore blurriness. We expect improved performance through an iterative multi-step transformation, such as a modern probabilistic diffusion generative model. Secondly, while we found some invariance to the input alignment, we would like to investigate whether the concatenated 3D inputs could be better aligned to match the target 3D locations and therefore support better modeling of some high-frequency details near their corresponding specific 3D slices. For example we could potentially optimize the input affine transformations, in particular their roto-translations, according to a secondary objective, repeated at inference as part of the modeling.

6. Conclusion

In conclusion, we found that simply repeating the 2D inputs into concatenated 3D volumes, then treating the 2D to 3D translation problem as a 3D to 3D translation task, leads to improved correlation between the synthesized outputs over more sophisticated multi-view latent integration approaches. While we expect other off-the-shelf 3D to 3D conditional generative modeling approaches to be immediately applicable within our framework, we found particular success by directly applying 3D to 3D neural optimal transport to attain highly correlated 3D synthesized outputs with their corresponding 2D inputs. The overall proposed approach is fast to train, data efficient, stable, and generalizes well to new out-of-distribution views making it applicable in a real-world clinical setting. However it does exhibit blurriness where there is uncertainty in the outputs; in the future it would be worth investigating iterative alignment optimization for the repeated 3D inputs as a secondary objective to the generative modeling, repeated during inference, potentially mitigating this uncertainty.

Acknowledgments This work was supported by CONAHCyT, and Durham University.

References

[1] S. G Armato III, Geoffrey McLennan, Luc Bidaut, Michael F McNitt-Gray, Charles R Meyer, Anthony P Reeves, Binsheng Zhao, Denise R Aberle, Claudia I Henschke, Eric A Hoffman, et al. The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans. *Medical physics*, 38(2):915–931, 2011. 5, 6, 7

[2] Samuel G Armato III, Lubomir Hadjiiski, Georgia D Tourassi, Karen Drukker, Maryellen L Giger, Feng Li, George Redmond, Keyvan Farahani, Justin S Kirby, and Laurence P Clarke. SPIE-AAPM-NCI Lung Nodule Classification Challenge Dataset. TCIA., 2015. 6

[3] Shaimaa Bakr, Olivier Gevaert, Sebastian Echegaray, Kelsey Ayers, Mu Zhou, Majid Shafiq, Hong Zheng, Jalen Anthony Benson, Weiruo Zhang, Ann NC Leung, et al. A radiogenomic dataset of non-small cell lung cancer. *Scientific data*, 5(1):1–9, 2018. 6

[4] Sam Bond-Taylor, Adam Leach, Yang Long, and Chris G Willcocks. Deep Generative Modelling: A Comparative Review of VAEs, GANs, Normalizing Flows, Energy-Based and Autoregressive Models. *IEEE TPAMI*, 44(11):7327–7347, 2021. 1, 3

[5] Abril Corona-Figueroa, Jonathan Frawley, Sam Bond-Taylor, Sarath Bethapudi, Hubert PH Shum, and Chris G Willcocks. MedNeRF: Medical Neural Radiance Fields for Reconstructing 3D-aware CT-Projections from a Single X-ray. In *2022 44th Annual Int. Conf. of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 3843–3848. IEEE, 2022. 1, 2, 5

[6] Abril Corona-Figueroa, Sam Bond-Taylor, Neelanjan Bhowmik, Yona Falinie A. Gaus, Toby P. Breckon, Hubert P. H. Shum, and Chris G. Willcocks. Unaligned 2D to 3D Translation with Conditional Vector-Quantized Code Diffusion using Transformers. In *Proc. of the IEEE/CVF Int. Conf. on Comput. Vis. (ICCV)*, pages 14585–14594, 2023. 2, 4, 5

[7] Qi Feng, Hubert PH Shum, and Shigeo Morishima. Enhancing Perception and Immersion in Pre-Captured Environments through Learning-Based Eye Height Adaptation. In *2023 IEEE Int. Symposium on Mixed and Augmented Reality (ISMAR)*, pages 405–414. IEEE, 2023. 1

[8] Jean Feydy. Geometric data analysis, beyond convolutions. *Applied Mathematics*, 2020. 3

[9] Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between Optimal Transport and MMD using Sinkhorn Divergences. In *The 22nd Int. Conf. on Artificial Intelligence and Statistics*, pages 2681–2690. PMLR, 2019. 4

[10] Vincent Fortuin. Priors in Bayesian Deep Learning: A Review. *Int. Statistical Review*, 90(3):563–591, 2022. 2

[11] Fabio Garcea, Alessio Serra, Fabrizio Lamberti, and Lia Morra. Data augmentation for medical imaging: A systematic literature review. *Comput.s in Biology and Medicine*, 152:106391, 2023. 2

[12] Aude Genevay, Gabriel Peyré, and Marco Cuturi. Learning generative models with sinkhorn divergences. In *Int. Conf. on Artificial Intelligence and Statistics*, pages 1608–1617. PMLR, 2018. 4

[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2

[14] Nathael Gozlan, Cyril Roberto, Paul-Marie Samson, and Prasad Tetali. Kantorovich duality for general transport costs and applications. *Journal of Functional Analysis*, 273(11): 3327–3405, 2017. 4

[15] Doga Gunduzalp, Batuhan Cengiz, Mehmet Ozan Unal, and Isa Yildirim. 3D U-NetR: Low Dose Computed Tomography Reconstruction via Deep Learning and 3 Dimensional Convolutions. *arXiv preprint arXiv:2105.14130*, 2021. 2

[16] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth, and Daguang Xu. Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images. In *Int. MICCAI Brainlesion Workshop*, pages 272–284. Springer, 2021. 3, 4

[17] Brian KS Isaac-Medina, Neelanjan Bhowmik, Chris G Willcocks, and Toby P Breckon. Cross-modal Image Synthesis within Dual-Energy X-ray Security Imagery. In *Proc. of the IEEE/CVF Conf. on Comput. Vis. and Pattern Recog.*, pages 333–341, 2022. 1

[18] Ling Jiang, Mengxi Zhang, Ran Wei, Bo Liu, Xiangzhi Bai, and Fugen Zhou. Reconstruction of 3D CT from A Single X-ray Projection View Using CVAE-GAN. In *2021 IEEE Int. Conf. on Medical Imaging Physics and Engineering (ICMIPE)*, pages 1–6, 2021. 2

[19] Hao Jin, Minghui Lian, Shicheng Qiu, Xuxu Han, Xizhi Zhao, Long Yang, Zhiyi Zhang, Haoran Xie, Kouichi Konno, and Shaojun Hu. A Semi-automatic Oriental Ink Painting Framework for Robotic Drawing from 3D Models. *IEEE Robotics and Automation Letters*, 2023. 1

[20] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *Adv. Neural Inform. Process. Syst.*, 33:12104–12114, 2020. 2

[21] Aghiles Kebaili, Jérôme Lapuyade-Lahorgue, and Su Ruan. Deep Learning Approaches for Data Augmentation in Medical Imaging: A Review. *J. of Imaging*, 9(4):81, 2023. 2

[22] Diederik P Kingma and Max Welling. Auto-Encoding Variational Bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2

[23] Alexander Korotin, Daniil Selikhanovych, and Evgeny Burnaev. Neural Optimal Transport. In *Int. Conf. on Learn. Represent.*, 2022. 7

[24] Daeun Kyung, Kyungmin Jo, Jaegul Choo, Joonseok Lee, and Edward Choi. Perspective Projection-Based 3d CT Reconstruction from Biplanar X-Rays. In *ICASSP 2023-2023 IEEE Int. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, pages 1–5. IEEE, 2023. 2

[25] Alex Junho Lee, Seungwon Song, Hyungtae Lim, Woojoo Lee, and Hyun Myung. $(LC)^2$: LiDAR-Camera Loop Constraints For Cross-Modal Place Recognition. *IEEE Robotics and Automation Letters*, 2023. 1

[26] Yanbin Liu, Girish Dwivedi, Farid Boussaid, Frank Sanfilippo, Makoto Yamada, and Mohammed Bennamoun. Inflating 2D convolution weights for efficient generation of

3D medical images. *Comput. Methods and Programs in Biomedicine*, 240:107685, 2023. 2

[27] Zhenyu Liu, Qide Wang, Daxin Liu, and Jianrong Tan. PA-Pose: Partial point cloud fusion based on reliable alignment for 6D pose tracking. *Pattern Recog.*, 148:110151, 2024. 1

[28] Kirsten WH Maas, Nicola Pezzotti, Amy JE Vermeer, Danny Ruijters, and Anna Vilanova. Nerf for 3d reconstruction from x-ray angiography: Possibilities and limitations. In *VCBM 2023: Eurographics Workshop on Visual Computing for Biology and Medicine*, pages 29–40. Eurographics Association, 2023. 2

[29] P Madhavi, S Patel, and AS Tsao. Data from Anti-PD-1 Immunotherapy Lung [Data set]. TCIA. *TCIA.*, 10, 2019. 6

[30] Payal Maken and Abhishek Gupta. 2D-to-3D: a review for computational 3D image reconstruction from X-ray images. *Archives of Computational Methods in Engineering*, 30(1): 85–114, 2023. 1

[31] Reza Moradi, Reza Berangi, and Behrouz Minaei. A survey of regularization strategies for deep models. *Artificial Intelligence Review*, 53:3947–3986, 2020. 2

[32] Youssef SG Nashed, Frederic Poitevin, Harshit Gupta, Geoffrey Woollard, Michael Kagan, Chuck Yoon, and Daniel Ratner. End-to-End Simultaneous Learning of Single-particle Orientation and 3D Map Reconstruction from Cryo-electron Microscopy Data. *arXiv preprint arXiv:2107.02958*, 2021. 2

[33] Sang Hyeon Oh, Kwak Dong Hwan, and Hyun Tek Lim. 2.5 D SLAM Algorithm with Novel Data Fusion Method Between 2D-Lidar and Camera. In *2023 23rd Int. Conf. on Control, Automation and Syst. (ICCAS)*, pages 1904–1907. IEEE, 2023. 1

[34] Pauliina Paavilainen, Saad Ullah Akram, and Juho Kannala. Bridging the gap between paired and unpaired medical image translation. In *MICCAI Workshop on Deep Generative Models*, pages 35–44. Springer, 2021. 2

[35] Md Aminur Rab Ratul, Kun Yuan, and WonSook Lee. CCX-rayNet: a class conditioned convolutional neural network for biplanar X-rays to CT volume. In *2021 IEEE 18th Int. Symposium on Biomedical Imaging (ISBI)*, pages 1655–1659. IEEE, 2021. 2, 5, 6, 8

[36] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Comput.-Assisted Intervention–MICCAI 2015: 18th Int. Conf., Munich, Germany, October 5-9, 2015, Proc., Part III 18*, pages 234–241. Springer, 2015. 4

[37] Joel Saltz, Mary Saltz, Prateek Prasanna, Richard Moffitt, Janos Hajagos, Erich Bremer, Joseph Balsamo, and Tahsin Kurc. Stony Brook University COVID-19 Positive Cases [Data set]. TCIA. *BBAG-2923*, 2021. 6, 7

[38] Hiroshi Sasaki, Chris G Willcocks, and Toby P Breckon. UNIT-DDPM: UNpaired Image Translation with Denoising Diffusion Probabilistic Models. *arXiv preprint arXiv:2104.05358*, 2021. 1

[39] Bowen Song, Liyue Shen, and Lei Xing. PINER: Prior-informed Implicit Neural Representation Learning for Test-time Adaptation in Sparse-view CT Reconstruction. In *Proc. of the IEEE/CVF Winter Conf. on Applications of Comput. Vis.*, pages 1928–1938, 2023. 2

[40] David Stojanovski, Uxio Hermida, Marica Muffoletto, Pablo Lamata, Arian Beqiri, and Alberto Gomez. Efficient Pix2Vox++ for 3D Cardiac Reconstruction from 2D echo views. In *Int. Workshop on Advances in Simplifying Medical Ultrasound*, pages 86–95. Springer, 2022. 2

[41] Yingjie Tian and Yuqi Zhang. A comprehensive survey on regularization strategies in machine learning. *Inform. Fusion*, 80:146–166, 2022. 2

[42] Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schoelkopf. Wasserstein auto-encoders. In *Int. Conf. on Learn. Represent.*, 2018. 3

[43] E Tsai, S Simpson, MP Lungren, M Hershman, L Roshkovan, E Colak, BJ Erickson, G Shih, A Stein, J Kalpathy-Cramer, et al. Medical Imaging Data Resource Center (MIDRC)-RSNA Int. COVID-19 Open Radiology Database (RICORD) Release 1b-Chest CT Covid-(MIDRC-RICORD-1B)[dataset]. *TCIA.*, 202010, 2020. 6, 7

[44] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *Proc. of the IEEE Conf. on Comput. Vis. and pattern recognition*, pages 9446–9454, 2018. 2

[45] Ge Wang, Jong Chul Ye, and Bruno De Man. Deep learning for tomographic image reconstruction. *Nature machine intelligence*, 2(12):737–748, 2020. 1

[46] Ethan Weinberger, Joseph Janizek, and Su-In Lee. Learning deep attribution priors based on prior knowledge. *Advances in Neural Inform. Process. Syst.*, 33:14034–14045, 2020. 2

[47] Zhisheng Xiao, Karsten Kreis, and Arash Vahdat. Tackling the generative learning trilemma with denoising diffusion GANs. In *Int. Conf. on Learn. Represent.*, 2022. 3

[48] Jinzhong Yang, Greg Sharp, Harini Veeraraghavan, Wouter Van Elmpt, Andre Dekker, Tim Lustberg, and Mark Gooding. Data from lung CT segmentation challenge. *TCIA.*, 2017. 6

[49] Xingde Ying, Heng Guo, Kai Ma, Jian Wu, Zhengxin Weng, and Yefeng Zheng. X2CT-GAN: reconstructing CT from biplanar X-rays with generative adversarial networks. In *Proc. of the IEEE/CVF Conf. on Comput. Vis. and Pattern Recog.*, pages 10619–10628, 2019. 2, 5, 6, 8

[50] Fangneng Zhan, Yingchen Yu, Rongliang Wu, Jiahui Zhang, Shijian Lu, Lingjie Liu, Adam Kortylewski, Christian Theobalt, and Eric Xing. Multimodal image synthesis and editing: A survey and taxonomy. *IEEE TPAMI*, 2023. 1

[51] Shengyu Zhao, Zhijian Liu, Ji Lin, Jun-Yan Zhu, and Song Han. Differentiable augmentation for data-efficient gan training. *Adv. Neural Inform. Process. Syst.*, 33:7559–7570, 2020. 8

[52] Jingyi Zuo. 2D to 3D Neurovascular Reconstruction from Biplane View via Deep Learning. In *Int. Conf. on Comput. and Data Sci.*, pages 383–387, 2021. 2