Unaligned 2D to 3D Translation with Conditional Vector-Quantized Code Diffusion using Transformers

Abril Corona-Figueroa[1] Sam Bond-Taylor[1] Neelanjan Bhowmik[1]
Yona Falinie A. Gaus[1] Toby P. Breckon[1,2] Hubert P. H. Shum[1] Chris G. Willcocks[1]
Department of {[1]Computer Science | [2]Engineering}, Durham University, Durham, UK
https://abrilcf.github.io/publications/CodeDiff3D

Abstract

Generating 3D images of complex objects conditionally from a few 2D views is a difficult synthesis problem, compounded by issues such as domain gap and geometric misalignment. For instance, a unified framework such as Generative Adversarial Networks cannot achieve this unless they explicitly define both a domain-invariant and geometric-invariant joint latent distribution, whereas Neural Radiance Fields are generally unable to handle both issues as they optimize at the pixel level. By contrast, we propose a simple and novel 2D to 3D synthesis approach based on conditional diffusion with vector-quantized codes. Operating in an information-rich code space enables high-resolution 3D synthesis via full-coverage attention across the views. Specifically, we generate the 3D codes (e.g. for CT images) conditional on previously generated 3D codes and the entire codebook of two 2D views (e.g. 2D X-rays). Qualitative and quantitative results demonstrate state-of-the-art performance over specialized methods across varied evaluation criteria, including fidelity metrics such as density, coverage, and distortion metrics for two complex volumetric imagery datasets from in real-world scenarios.

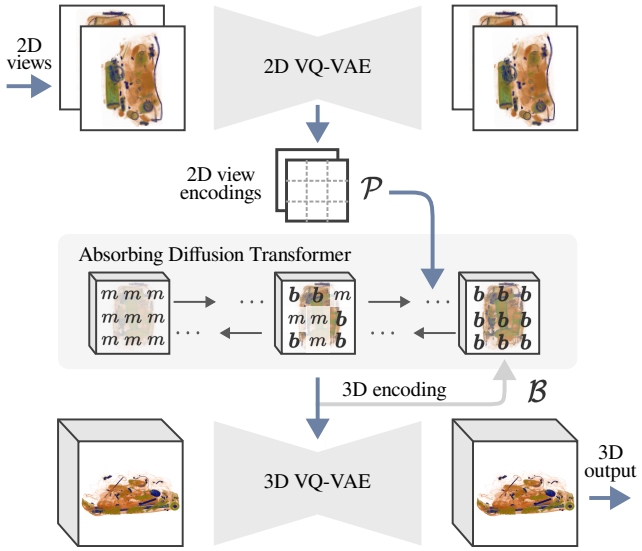

Figure 1. The proposed approach independently learns a discrete information-rich codebook for 2D and 3D domains with two VQ-VAEs. New 3D samples are then synthesized conditional on the complete codebook of multiple 2D views with a transformer.

1. Introduction

3D imaging is essential in several fields, from clinical applications with Magnetic Resonance Imaging (MRI) and Computed Tomography (CT) modalities [13, 48], to virtual/augmented reality [8,27], self-driving vehicles [23,31], and security [42, 45]. However, the diverse range of available imaging devices exhibit differences in cost, quality, and accessibility, which has led to an increased interest in leveraging 2D imaging for 3D acquisition. For instance, in hospitals, CT from biplanar X-rays could minimize patient exposure to a substantial dose of ionizing radiation [12]. Similarly, at airports, volumetric reconstruction from security baggage screening could be more effective at detecting prohibited items [43,44]. The capability to seamlessly translate between these imaging domains, eliminating their limitations, is a high-impact application; however, this remains a difficult conditional generative modeling problem.

Generating accurate 3D representations from 2D images poses a significant challenge in the field of computer vision. Approaches must account for complex shape topologies, fine-grained textures, domain differences, and incomplete input information. Current methods aim to solve this under-constrained problem by relying on cycle consistency losses [50] for disentangling 3D shape properties like appearance and viewpoint from a single input image [3], improving fine-grained 3D mesh attributes [21], for 3D-aware face image generation by distilling StyleGAN2 latent space [32] and recovering 3D shapes from videos [47]. Despite

these models being unpaired, they require separate convolutional neural network (CNN) designs for predicting each 3D property, and they cannot generate accurate 3D representations without additional 2D supervisory signals.

In this work, we present a unified approach that tackles 3D synthesis by reformulating it as a voxel-to-voxel prediction problem. We achieve this by conditioning an unconstrained transformer on 2D input views. While some recent methods have used transformers for multi-view 3D object reconstruction, they suffer from computational inefficiency and are sensitive to domain shifts and input view irregularities [29, 41].

We investigate 2D to 3D image translation of complex data that exhibit varying outer and internal topologies with different density properties and domains. To achieve this, we propose a novel two-stage translation approach that models the conditional probability of generating a 3D image given 2D input views with a discrete diffusion model. By applying this process in a highly compressed discrete latent space, our approach can extract high level features of complex objects and scale to high-resolution data without requiring paired datasets (Fig. 1), which is a desired feature for real-world applications. First, our approach learns information-rich discrete spaces of 2D and 3D distributions independently with two Vector-Quantized Variational Autoencoders (VQ-VAE), removing the need for alignment of both geometries. Second, we use a diffusion model parameterized by an unconstrained transformer that allows bidirectional 2D global context when generating the 3D representation, improving feature learning and speeding up the sampling process.

To summarize, our main contributions are:

- We propose a novel and simple translation approach based on conditional diffusion using transformers, generating high-quality 3D samples conditional on two 2D images from a different domain.

- We show that diffusion in the information-rich latent code space not only allows for our model to scale easily to high-resolution, but also allows for translation of unaligned inputs—as our full-coverage attention on latent encodings permits any part of all 2D inputs to contribute to voxel predictions.

- The model is shown to give state-of-the-art density and coverage over competing methods such as Generative Adversarial Networks (GAN) and Neural Radiance Fields (NeRF) while offering true likelihood estimates.

2. Prior Work

2.1. Autoregressive Modeling

Autoregressive models are a class of likelihood-based generative models that have been demonstrated to be potent density estimators, exhibiting greater training stability and generalization capabilities [30, 37] compared to implicit generative models such as GAN [18]. They break down the joint distribution of structured outputs into products of conditional distributions,

$$p(\boldsymbol{c}|\boldsymbol{Z}) = \prod_{i=1}^{L} p_\theta(c_i|c_1, ..., c_{i-1}; \boldsymbol{Z}). \tag{1}$$

However, as their receptive field is limited to previously generated tokens, their representation ability is biased, and images do not conform to such sequential manner. Moreover, this also restricts them to relatively low dimensional data [36, 38].

2.2. Vector-Quantized Representations

The vector quantization (VQ) technique has been adopted by explicit generative models to alleviate issues including scaling, posterior collapse and blurred outputs by quantizing the latent representations to a fixed number $\{\boldsymbol{q}^1, \cdots, \boldsymbol{q}^K\}$ [19, 39, 49]. Furthermore, VQ-based models have achieved sharper reconstructions than implicit generative models on continuous latent variables. Following their success in generative modeling, we make use of VQ representations, where a convolutional autoencoder extracts high-level features to an information-rich latent space. VQ image models [39], which compress images to a low dimensional discrete latent space and subsequently model this space with a powerful generative model, have recently been used for a variety of tasks. Chen et al., [10] address the domain gap issue in cross-domain analysis by introducing VQ into the image-to-image translation framework; however, their approach deals with the two data modalities having the same dimension and relies on spatial correlations.

2.3. 2D to 3D Image Translation

Image-to-image translation methods which aim to reconstruct a 3D representation given a single or multiple 2D images, have achieved notable success within the field [16]. Generally, these architectures first extract 2D features from the input image into a latent vector, which can be fused with other information such as geometric priors, and lastly, the decoder generates the predicted 3D representation [16, 20, 33]. However, these techniques applied to natural images do not easily translate to real-application domains. Moreover, most conditional generative models have dealt with the input and output data having the same dimension, i.e. 2D to 2D or 3D to 3D.

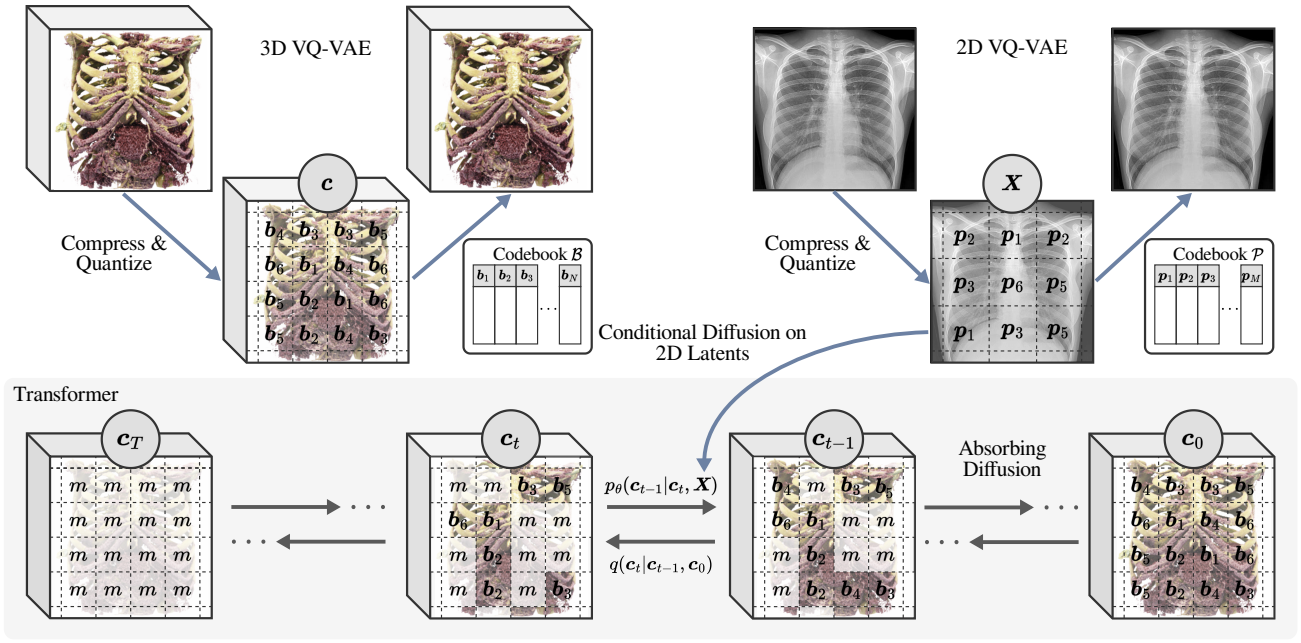

Figure 2. Our approach allows 2D to 3D translation from unaligned data by first compressing to an information rich Vector-Quantized discrete space, then modeling the conditional probability with a discrete diffusion model parameterized by an unconstrained transformer.

Generative Adversarial Networks. GAN [18] have emerged as representatives of implicit generative modeling due to their capability to generate realistic synthesized images. However, they often suffer from training instability [1]. In our domain's context, we are comparing two GAN-based models. The first is X2CT-GAN [48], which utilizes a 3D GAN with skip-connections, along with an additional CycleGAN for capturing style differences. The second model is CCX-rayNet [28], which employs a class-conditioned module to reconstruct a CT from 2D X-ray images. Unlike CNN-based models that require aligned inputs (i.e. consistent resolution, orientation, and transformations) and often rely on skip-connections, our approach allows unaligned inputs via full-coverage attention on an information-rich discrete latent space.

Neural Radiance Fields. The ability of NeRF to generate novel views of complex 3D scenes from a partial set of 2D images [25], inspired MedNeRF [13] for rendering CT projections from a small set or single-view X-rays. While NeRF render 3D-informed images from 2D viewpoints, they differ fundamentally from our method. In particular, they are focused on object and scene representation conditional on coordinate information, rather than novel synthesis and generalizability [17]. In contrast, our approach needs no prior data on camera positions or aligned inputs. It enables direct sampling based on input views while providing control over the generative process.

2.4. Hybrid Generative Models

A way to address issues in specific generative models such as long training or poor scaling is by combining two or more approaches [7]. For instance, transformers and their self-attention mechanism use in an encoder-decoder setup [40] to improve both autoregressive models and other generative models due to their stable training and ability to learn long-distance dependencies. This is achieved using the attention scheme,

$$\text{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_k}}\right)\boldsymbol{V}. \quad (2)$$

where the values $\boldsymbol{V}$ represent the encoded inputs, the keys $\boldsymbol{K}$ are used for indexing and queries $\boldsymbol{Q}$ to select specific values. Additionally positional information is passed into the function via fixed or trainable positional embeddings [11].

Of particular interest to this work is the two-stage process proposed by Bond-Taylor *et al.* [6], where the discrete latent space is modeled by a diffusion model parameterized by an unconstrained transformer that learns to unmask the data. This process allows faster sampling because multiple tokens can be predicted in parallel. Inspired by this work, we bridge the domain gap between 2D to 3D translation by combining the infinite receptive fields of attention for representing both data distributions and allowing learning complex topologies with the powerful feature-extraction ability and scalability of VQ-VAE. One of the advantages of our approach is that it does not require spatially aligned 2D and

3D samples, thus avoiding issues with both the geometric and domain misalignment between the two modalities.

3. Method

In this work, we address the task of synthesizing complex 3D data portraying varying topologies and material properties (e.g. CT-like images) given a two 2D views taken from different angles (e.g. X-rays). In particular, we avoid dependencies on both the geometric and domain relationships between the 2D and 3D data which are significant sources of real-world data misalignment caused by object/device movement and characteristics. In this setting, learning a deterministic mapping between modalities is impractical due to a large number of possible input/output pairings in real-world data, meaning that outputs would be very blurry and unhelpful [35] to human operators for instance. As such, we propose modeling the mapping between 2D and 3D data with a conditional likelihood-based generative model, allowing sampled 3D data to sit at arbitrary positions/rotations relative to the 2D data.

Formally, given 3D data $\boldsymbol{I} \in \mathbb{R}^{H\times W\times D}$, and multiple corresponding 2D views $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}, \boldsymbol{x}_i \in \mathbb{R}^{H\times W}$, we wish to learn the conditional distribution $p(\boldsymbol{I}|\boldsymbol{X})$. Training a transformer model directly on complex high-dimensional pixel data would be impractical, as the computational complexity would increase quadratically from the self-attention mechanism [34, 40]. To overcome this, we decompose the task into two stages to take advantage of the power and flexibility of transformers. Stage 1 separately learns to compress the 2D and 3D data to small but information-rich discrete latent spaces that accurately represent the data. In Stage 2, we model the conditional probability distribution in the learned compressed discrete space with a discrete diffusion model parameterized by a powerful unconstrained transformer that is spatially invariant over the 2D data. The overall process is illustrated in Fig. 2, and train it alternatively with a patch-wise discriminator that uses a combination of spatial and style augmentations for both 2D and 3D images. More details in Appendix C.

3.1. Stage 1: Unpaired Compression

Separately for 2D and 3D data, we learn to compress a single data point $\boldsymbol{x}$ (each 2D view is also compressed separately) to a relatively small set of integer values $\boldsymbol{z}$. This is performed using a VQ-VAE [39], which achieves extremely high compression rates by utilising a codebook for information rich vectors, each of which is able to represent an image patch, while a neural decoder models how these codes interact.

The VQ-VAE approach is of particular interest to our work as it allow a 2-stage scheme where a compact and quantized latent space can be learned through a convolutional autoencoder with a large receptive field. In addition, variational autoencoders (VAE) [22] allow us to provision per-3D-image likelihood estimates, in contrast to other 2D-3D translation models based on GAN, which cannot directly provide a probability interpretation.

In particular, a convolutional encoder downsamples data $\boldsymbol{x}$ to a smaller spatial resolution, $E(\boldsymbol{x}) = \{\boldsymbol{e}_1, \boldsymbol{e}_2, ..., \boldsymbol{e}_L\} \in \mathbb{R}^{L\times D}$. Each continuous encoding $\boldsymbol{e}_i$ is subsequently quantized by mapping to the closest element in the codebook of vectors $\mathcal{B} \in \mathbb{R}^{K\times D}$, where K is the number of discrete codes in the codebook, and D is the dimension of each code,

$$\boldsymbol{z}_q = \{\boldsymbol{q}_1, \boldsymbol{q}_2, ..., \boldsymbol{q}_L\} \text{ , where } \boldsymbol{q}_i = \min_{\boldsymbol{b}_j\in\mathcal{B}} \|\boldsymbol{e}_i - \boldsymbol{b}_j\|, \quad (3)$$

with the straight-through gradient estimator [5] used to approximate the gradients through this non-differentiable process. The discrete latents are then decompressed with a convolutional decoder $\hat{\boldsymbol{x}} = G(\boldsymbol{z}_q)$. The model is trained end-to-end by minimizing the loss,

$$\mathcal{L}_{\text{VQ}} = \mathcal{L}_{\text{rec}}(\boldsymbol{x}, \hat{\boldsymbol{x}}) + \|\text{sg}[E(\boldsymbol{x})] - \boldsymbol{z}_q\|_2^2 + \beta\|\text{sg}[\boldsymbol{z}_q] - E(\boldsymbol{x})\|_2^2. \quad (4)$$

which balances reconstruction quality $\mathcal{L}_{\text{rec}}$ against quantization terms that encourage the encoding $E(\boldsymbol{x})$ to match the closest codes in the codebook.

Our approach eliminates the need for spatially aligned 2D and 3D images, which in practice are difficult to obtain accurately. In addition, the flexibility of our model allows the use of an arbitrary number of 2D input views without the requirement of camera priors or changing the network architecture as this can be effectively achieved by using smaller or larger latent code sizes. An ablation of these aspects can be found in the Appendix A.

3.2. Stage 2: Conditional Discrete Diffusion

To translate data from 2D to 3D we model the conditional probability distribution of 3D data given a few 2D views using a discrete diffusion model in the learned vector-quantized space, $p(\boldsymbol{c}|\boldsymbol{X})$, where $\boldsymbol{c}$ represents the VQ codes of the 3D data and $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ represents the set of VQ codes of the all the 2D views. Specifically, we use the discrete absorbing diffusion process [4, 6], which is more suitable for this task due to its bidirectional nature, allowing them to outperform autoregressive approaches while also being able to scale to higher dimensional spaces, faster sampling, and being substantially less prone to overfitting [9] which is crucial when training on small datasets (as is generally the case for real-world datasets). More detail on this aspect is presented in the Appendix A (Fig. 4).

In this case, the discrete 3D latents are gradually masked out using randomized orders over a number of time steps T such that at time step t, $\boldsymbol{c}^t$ is defined as the discrete 3D latent $\boldsymbol{c}$ with each token masked out with probability $\frac{t}{T}$. Formally,

this masking procedure is defined by a Markov chain,

$$q(\boldsymbol{c}^{1:T}|\boldsymbol{c}^0) = \prod_{t=1}^{T} q(\boldsymbol{c}^t|\boldsymbol{c}^{t-1}). \quad (5)$$

The posterior is defined as $q(\boldsymbol{c}_t|\boldsymbol{c}_{t-1}) = \text{cat}(\boldsymbol{c}_t; \boldsymbol{p} = \boldsymbol{c}_{t-1}\boldsymbol{Q}_t)$ for one-hot $\boldsymbol{c}_{t-1}$ where $\boldsymbol{Q}_t = (1-\beta_t)I + \beta_t \mathbb{1} e_m^T$ is a matrix denoting the probabilities of moving to each successive state, e_m is a vector with a one on mask states m and zeros elsewhere, and $\beta_t = \frac{1}{T-t+1}$.

The reverse of this diffusion process is another Markov chain that gradually unmasks the latents

$$p_\theta(\boldsymbol{c}^{0:T}|\boldsymbol{Z}) = \prod_{t=1}^{T} p_\theta(\boldsymbol{c}^{t-1}|\boldsymbol{c}^t, \boldsymbol{X}). \quad (6)$$

This can be approximated by training an unconstrained transformer to predict the original latents from the masked ones, optimizing the Evidence Lower Bound (ELBO),

$$\mathbb{E}_{q(\boldsymbol{c}^0,\boldsymbol{X})} \sum_{t=1}^{T} \gamma\, \mathbb{E}_{q(\boldsymbol{c}^t|\boldsymbol{c}^0)} \Big[\sum_{[\boldsymbol{c}^t]_i = m} \log p_\theta([\boldsymbol{c}^0]_i|\boldsymbol{c}^t, \boldsymbol{X}) \Big], \quad (7)$$

where $\gamma = \frac{T-t+1}{T}$ is a reweighting term used to improve convergence [6].

Discrete vs. Continuous Latents Discrete representations are important for our approach because applying self-attention on a compact, information-rich space is more efficient than attending over a larger, continuous space due to limitations in sequence length [34] coupled with less effective integration of information [24]. Moreover, compressed representations have been shown to improve generalization without relying on complex hierarchical architectures [46].

Full-coverage Attention To leverage the compressed latent space from the diffusion model, we employ an unconstrained transformer to parameterize the denoising function. By flattening and concatenating 2D latents, we integrate discrete representations into a 1D manifold, aided by a trainable positional embedding. In contrast to CNN reliance on local alignment and/or deep architectures to increase their limited receptive fields, our method provides global context through information-rich discrete representations, allowing all parts of the 2D inputs to influence voxel predictions [24].

Domain Invariance Our approach is robust to differences in the distributions of 2D and 3D data domains as we model distributions over discrete latent representations separately rather than low-level voxels in a joint manner. In contrast, a unified framework such as GAN cannot achieve this unless they explicitly define a domain-invariant joint latent distribution. Domain invariance is of particular interest when dealing with real-world scenarios where data exhibits complex variations, such as different camera perspectives, lighting conditions, or imaging modalities. Examples of this are presented in the Appendix A (Fig. 4).

Likelihood Estimation Due to the fact that the VQ-VAE decoder is trained with MAP-inference, within this framework we are able to estimate the conditional log likelihood $\log p(\boldsymbol{I}|\boldsymbol{X}) \approx \log p(\boldsymbol{I}|\boldsymbol{c})p(\boldsymbol{c}|\boldsymbol{X})$ [39], where $p(\boldsymbol{I}|\boldsymbol{c})$ is the 3D VQ-VAE decoder, and $p(\boldsymbol{c}|\boldsymbol{X})$ is the conditional discrete diffusion model. Subsequently, our model provisions per-3D-image likelihood estimates, which provide a distance measurement from the true distribution of the data.

4. Experiments

This section evaluates the ability of our unconstrained transformer to perform 2D to 3D translation using discrete VQ representations against the SOTA models on this task: X2CT-GAN [48] and CCX-rayNet [28] on two datasets complex volumetric imagery from the publicly available chest CT scans, LIDC-IDRI [2], and security baggage screening. Additionally, we provide a 2D evaluation compared to the MedNeRF [13] model in terms of MIPs in Sec. 4.1.3. We evaluate in terms of negative log-likelihood (NLL), Density and Coverage, and distortion metrics, including SSIM, PSNR, MSE and MAE. The best values are in bold, and the second-best values are underlined.

4.1. Conditional 3D Modeling on 2D Views

2D to 3D translation is performed by our unconstrained transformer trained on the discrete latents encodings computed by the two VQ-VAE. At inference, the 3D latents are gradually predicted using the conditioning information from two 2D inputs in the form of 16×16 codes, which are concatenated at the start.

Fidelity Metrics. We aim to measure the quality as well as the overlap between the manifold of generated samples and the manifold of real data. One of our primary motivations for using likelihood-based generative models is that Negative Log-Likelihood (NLL) allows us to monitor over-fitting effectively. This is a potential issue when training models on real application datasets which are generally much smaller than natural image datasets. In addition, we report values based on Density-Coverage [26], which independently assesses the fidelity and variability of a model. In these metrics, images are first projected into an embedding space, and a scoring function estimates the manifold density in the neighborhood of each embedded data point. Generative models trained with natural images mostly rely on the features from an ImageNet pretrained model for evaluation. Since our target data is distinct from ImageNet samples, to compute these metrics, we use the features of a randomly

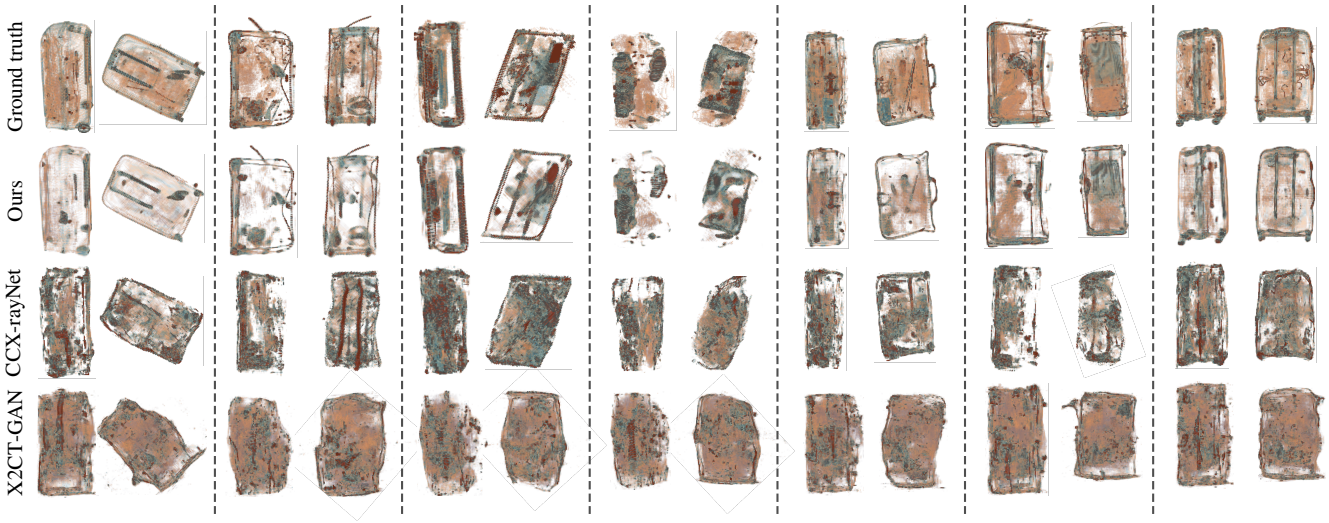

Figure 3. Comparison of 3D CT model samples trained on the baggage screening dataset, showing the ground truth, our method, CCX-rayNet [28] and X2CT-GAN [48].

Baggage Security Screening dataset							
Method	↓ NLL	↑ Density	↑ Coverage	↑ SSIM	↑ PSNR	↓ MSE	↓ MAE
X2CT-GAN	N/A	0.95	0.80	0.655	34.68	0.0014	0.0129
CCX-rayNet	N/A	1.28	0.89	0.886	35.45	0.0012	0.0069
Ours	**0.007**	**2.01**	**0.97**	**0.899**	**39.45**	**0.0007**	**0.0049**

Table 1. **3D fidelity and distortion metrics on the Baggage security dataset.** We compare our method and SOTA models in terms of fidelity and diversity (density and coverage). Additionally we evaluate the quality of generated voxel grids using distortion metrics (SSIM, PSNR, MSE and MAE).

initialized convolutional encoder architecture, as proposed in the work of [26]. Specifically, we obtained embeddings of dimension 2,197 of the real validation data points and samples from our model and competing models.

Distortion Metrics. We emphasize that commonly used distortion metrics such as Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) correlate poorly with human visual perception, as they ignore global structure and only focus on signal fidelity [14]. A model that simply optimizes pixel-level distortions (e.g. small adjustments in brightness, saturation etc.) can obtain a high score on these metrics. However, we also report values on SSIM, PSNR, Mean Squared Error (MSE) and Mean Absolute Error (MAE) for the research community as they are widely used for comparison [28, 48].

4.1.1 Modeling Objects in Baggage Screening

We evaluate our approach on a set of *"stream of commerce"* CT volumes from an aviation security context. The total dataset consists of 5,964 CT bag volumes. We resample the bag volume to $1 \times 1 \times 1$ mm^3, and subsequently, we resize it to a $256 \times 256 \times 256$ mm^3 cubic area. This dataset comprises authentic bags and suitcases, depicting both the common and the forbidden items during air travel. We generated a 2D subset of eight views, each taken at 45-degree azimuth intervals per bag. For our model, we set $|\boldsymbol{b}_i| = 256$, a codebook of 4,096 and train our transformer to predict sequences of length $16 \times 16 \times 16$.

Results. Table 1 reports results for conditional 3D modeling on two 2D input views comparing samples from our model and reconstructions from the competing approaches [48] and [28] for the baggage dataset. This dataset depicts varying topologies as there exists a wide variety of baggage, such as suitcases, and backpacks, in different sizes, styles, and materials. Moreover, the items/objects within passenger bags can be unpredictable and their arrangement and level of compactness could complicate their identification. Despite the challenging complexity of this dataset, we find that our approach offers superior performance, with a very large margin in terms of density ($+111\%$ improvement over X2CT-GAN [48], and $+53\%$ over CCX-rayNet [28]) and coverage ($+21\%$ over X2CT-GAN and $+8\%$ over CCX-rayNet). In terms of distortion metrics, our samples are of higher quality, with a significant improvement in PSNR, MSE and MAE compared to the competing models [28, 48].

Figure 4. Comparison of 3D CT model samples trained on the medical chest (LIDC-IDRI dataset [2]) from different anatomical planes (coronal and sagittal), showing the ground truth, our method, CCX-rayNet [28] and X2CT-GAN [48].

LIDC-IDRI (chest) dataset							
Method	↓ NLL	↑ Density	↑ Coverage	↑ SSIM	↑ PSNR	↓ MSE	↓ MAE
X2CT-GAN	N/A	0.87	0.88	0.321	19.68	0.045	0.151
CCX-rayNet	N/A	1.41	**0.98**	0.386	22.66	**0.032**	0.108
Ours	**0.10**	**1.42**	0.97	**0.436**	**25.05**	0.048	**0.013**

Table 2. **3D fidelity and distortion metrics on LIDC-IDRI (chest) dataset.** We followed the experimental protocol X2CT-GAN [48] and CCX-rayNet [28] with a wider Hounsfield unit range of -1,000 HU to +1,000 HU. We show additional analysis on out-of-distribution inputs in the Appendix A (Fig. 4).

In Fig. 3 we present a set of 3D generated samples from our approach and 3D reconstructions of the other models for the baggage dataset. It can be observed that the identity of each of the bags and suitcases is accurately modeled by our model, while CCX-rayNet [28] is limited at predicting the shape boundaries, and X2CT-GAN [48] shows a lack of diversity. It is worth noting that airport security officials focus on denser objects as these are more likely to be prohibited compared to less dense items such as clothes. In this context, our model is better able to model latent representations of denser structures, as indicated by darker colors, which aligns with real-world scenarios. In contrast, the other models depict blurring for both dense and soft structures, making it challenging to distinguish them.

4.1.2 Modeling Chest Anatomical Structures

We conduct a set of experiments using the publicly available dataset of chest CT scans, LIDC-IDRI [2]. We generated the corresponding X-ray projections using digitally reconstructed radiograph technology (DRR). For our model we set $|\boldsymbol{b}_i| = 256$, a codebook of 1,024 and train our transformer to predict sequences of length $8 \times 8 \times 8$.

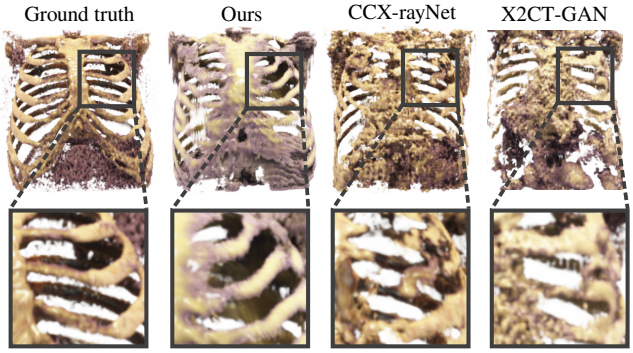

Figure 5. Local-level quality comparison of chest CT samples.

Results. As presented in Table 2, our method exhibits superior performance across almost all categories for density and coverage and surpasses the other models in all distortion metrics, achieving a 12% improvement in SSIM, 11% in PSNR, 41% in MSE, and 28% for MSE over the most competitive model, CCX-rayNet [28]. Furthermore, in comparison to X2CT-GAN [48], our model demonstrates an even greater degree of improvement performance.

In Fig. 4 we present a set of 3D generated samples from our approach and 3D reconstructions of the other models for the LIDC (chest) dataset. Although the sizes of both

datasets are relatively small (compared to natural image datasets), our model displays no signs of overfitting and can produce diverse shapes that closely align with the ground truth. In particular, our approach achieves a greater degree of precision in modeling and disentangling soft tissue and bone structures. This is evidenced by the distinctive colors employed in the renderings, with a purple*ish* tone denoting soft tissue and a beige color representing bone. By contrast, the other methods produce blurred images where anatomical structures overlap, and as a result, densities cannot be accurately modeled since they are all merged together. For instance, samples from our approach are less noisy, while the other models are only able to reasonably predict some parts of the volume and result in spatial blur/uncertainty in others (Fig. 5).

4.1.3 Analysis on Maximum Intensity Projections

We perform an additional evaluation in terms of Maximum Intensity Projections (MIP) in 2D, as they provide a representation of the highest intensity values along the spatial dimensions, accentuating low-level structures, which might be difficult to detect on individual slices of the 3D data. MIP are frequently used in real-world settings for clinical diagnosis, planning, object identification, or to simply enhance visibility. For our case, this could provide an additional interpretation of both quantitative and qualitative results.

In addition to X2CT-GAN [48], and CCX-rayNet [28], we also compare our model with the NeRF-based method MedNeRF [13], which has been designed to handle complex images such as X-rays. For this, we train MedNeRF [13] on both datasets and test them to render 3D-aware CT projections. For the other models, we obtained the MIP of their spatial dimensions and calculated density and coverage, SSIM and PSNR. These results can be found in Table 3. Qualitative results of this comparison are presented in the Appendix B.

4.2. Implementation Details

The 2D and 3D VQ-VAE from the first stage were trained simultaneously on independent NVIDIA A100 cards with a batch size of 32 and 8, respectively. For the VQ-AE we use the framework proposed in the work of [15], which substantially improves compression rates by learning a more information-rich codebook, while still permitting likelihood estimates. The 2D model takes approximately half the time of the 3D model to complete 100,000 iterations. Our unconstrained transformer from the second stage takes less than 5 hours to also complete 100,000 iterations; it effectively models 3D tokens very quickly. We use weights from 500k iterations for our reconstructions (Table 4). The transformer can easily fit into memory on a GPU with 12GB of VRAM with a batch size of 20 while the VQ-VAEs can also fit in such a GPU. In practice, this requires the use of small batch sizes making training in reasonable times less practical.

(a) Baggage Security Screening dataset				
Method	↑ D	↑ C	↑ SSIM	↑ PSNR
X2CT-GAN [48]	0.51	0.68	0.65	34.49
CCX-rayNet [28]	0.96	0.95	0.88	35.49
MedNeRF [13]	0.91	0.46	0.79	25.11
Ours	**1.84**	**0.99**	**0.91**	**39.43**
(b) LIDC-IDRI (Chest) dataset				
X2CT-GAN [48]	0.96	0.85	0.32	21.38
CCX-rayNet [28]	0.76	0.87	0.40	24.20
MedNeRF [13]	0.90	0.80	0.38	**27.02**
Ours	**1.17**	**0.91**	**0.42**	25.25

Table 3. **2D (Maximum Intensity Projections) fidelity and distortion metrics on both datasets.** We compare MedNeRF [13] in addition to SOTA models and our approach.

Component	Training time (hs)	Inference time (s)
2D VQ-VAE	~ 23	~ 0.01
3D VQ-VAE	~ 56	~ 0.27
Transformer	~ 4.5	$\lesssim 10.9$

Table 4. Training times (each for 100k iterations) and inference times for our approach, all performed on single NVIDIA A100 GPUs. The VQ-VAEs are independent and can be trained simultaneously on different GPUs.

5. Conclusion

We propose a novel 2D to 3D translation approach based on conditional diffusion using transformers. We find that compressing each domain independently offers several key advantages. The discrete compressed space allows for both fast and high-resolution image synthesis, where the 2D and 3D compression networks can be trained and verified independently without requiring aligned datasets. In particular, full-coverage attention over the complete information-rich 2D codebooks from a few views significantly improves the synthesis of new 3D images, where any part of all 2D inputs can contribute to the voxel predictions.

The proposed approach is surprisingly simple and intuitive in practice as diffusion models are shown to have excellent mode coverage giving diverse samples [9]. In the future, we would like to consider scaling our approach with larger models trained on more diverse datasets, provide an in-depth study on hallucinated outputs, and see how well it can generalize between very different imaging modalities.

Acknowledgments This work was supported by CONAHCyT, and by the EPSRC NorthFutures project (ref: EP/X031012/1).

References

[1] Martin Arjovsky and Léon Bottou. Towards Principled Methods for Training Generative Adversarial Networks. *arXiv preprint arXiv:1701.04862*, 2017. 3

[2] 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, 7

[3] Tristan Aumentado-Armstrong, Alex Levinshtein, Stavros Tsogkas, Konstantinos G. Derpanis, and Allan D. Jepson. Cycle-Consistent Generative Rendering for 2D-3D Modality Translation. In *2020 Int. Conf. on 3D Vision*, 2020. 1

[4] Jacob Austin, Daniel Johnson, Jonathan Ho, Danny Tarlow, and Rianne van den Berg. Structured Denoising Diffusion Models in Discrete State-Spaces. *arXiv preprint arXiv:2107.03006*, 2021. 4

[5] Yoshua Bengio. Estimating or propagating gradients through stochastic neurons. *arXiv preprint arXiv:1305.2982*, 2013. 4

[6] Sam Bond-Taylor, Peter Hessey, Hiroshi Sasaki, Toby P Breckon, and Chris G Willcocks. Unleashing Transformers: Parallel Token Prediction with Discrete Absorbing Diffusion for Dast High-Resolution Image Generation from Vector-Quantized Codes. In *European Conf. on Comput. Vis.*, pages 170–188. Springer, 2022. 3, 4, 5

[7] 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 Trans. Pattern Anal. Mach. Intell.*, 2021. 3

[8] Ziyi Chang, George A. Koulieris, and Hubert P. H. Shum. 3D Reconstruction of Sculptures from Single Images via Unsupervised Domain Adaptation on Implicit Models. In *Proc. of the 28th ACM Symposium on Virtual Reality Software and Technology*, VRST '22, New York, NY, USA, 2022. Association for Computing Mach.ry. 1

[9] Ziyi Chang, George A. Koulieris, and Hubert P. H. Shum. On the Design Fundamentals of Diffusion Models: A Survey. *arXiv preprint arXiv:2306.04542*, 2023. 4, 8

[10] Yu-Jie Chen, Shin-I Cheng, Wei-Chen Chiu, Hung-Yu Tseng, and Hsin-Ying Lee. Vector Quantized Image-to-Image Translation. In *European Conf. on Comput. Vis.*, pages 440–456. Springer, 2022. 2

[11] Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating Long Sequences with Sparse Transformers. *arXiv preprint arXiv:1904.10509*, 2019. 3

[12] Mary Coffey and Aude Vaandering. Patient Setup for PET/CT Acquisition in Radiotherapy Planning. *Radiotherapy and Oncology*, 96(3), 2010. PET in Radiotherapy Planning. 1

[13] Abril Corona-Figueroa, Jonathan Frawley, Sam Bond Taylor, Sarath Bethapudi, Hubert P. H. 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)*, 2022. 1, 3, 5, 8

[14] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P. Simoncelli. Comparison of Full-Reference Image Quality Models for Optimization of Image Processing Systems. *Int. Journal of Comput. Vis.*, 129(4):1258–1281, Apr. 2021. 6

[15] Patrick Esser, Robin Rombach, and Björn Ommer. Taming Transformers for High-Resolution Image Synthesis. *arXiv:2012.09841*, 2021. 8

[16] Kui Fu, Jiansheng Peng, Qiwen He, and Hanxiao Zhang. Single Image 3D Object Reconstruction based on Deep Learning: A review. *Multimedia Tools and Applications*, 80, 2021. 2

[17] Kyle Gao, Yina Gao, Hongjie He, Denning Lu, Linlin Xu, and Jonathan Li. NeRF: Neural Radiance Field in 3D Vision, A Comprehensive Review. *arXiv preprint arXiv:2210.00379*, 2022. 3

[18] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Networks. *Commun. ACM*, 63(11):139–144, oct 2020. 2, 3

[19] Ligong Han, Jian Ren, Hsin-Ying Lee, Francesco Barbieri, Kyle Olszewski, Shervin Minaee, Dimitris Metaxas, and Sergey Tulyakov. Show Me What and Tell Me How: Video Synthesis via Multimodal Conditioning. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, June 2022. 2

[20] Xian-Feng Han, Hamid Laga, and Mohammed Bennamoun. Image-Based 3D Object Reconstruction: State-of-the-Art and Trends in the Deep Learning Era. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(5):1578–1604, 2021. 2

[21] Tao Hu, Liwei Wang, Xiaogang Xu, Shu Liu, and Jiaya Jia. Self-Supervised 3D Mesh Reconstruction From Single Images. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pages 6002–6011, June 2021. 1

[22] Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In *Int. Conf. Mach. Learn.*, 2014. 4

[23] Li Li, Khalid N. Ismail, Hubert P. H. Shum, and Toby P. Breckon. DurLAR: A High-fidelity 128-channel LiDAR Dataset with Panoramic Ambient and Reflectivity Imagery for Multi-modal Autonomous Driving Applications. In *2021 Int. Conf. on 3D Vision*, 2021. 1

[24] Chengzhi Mao, Lu Jiang, Mostafa Dehghani, Carl Vondrick, Rahul Sukthankar, and Irfan Essa. Discrete Representations Strengthen Vision Transformer Robustness. In *Int. Conf. on Learning Representations*, 2022. 5

[25] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. *Communications of the ACM*, 65(1), 2021. 3

[26] Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable Fidelity and Diversity Metrics for Generative Models. In Hal Daumé III and Aarti Singh, editors, *Proc. of the 37th Int. Conf. on Mach. Learning*, volume 119 of *Proc. of Mach. Learning Research*, pages 7176–7185. PMLR, 13–18 Jul 2020. 5, 6

[27] Jae-Hyeung Park and Byoungho Lee. Holographic Techniques for Augmented Reality and Virtual Reality Near-Eye Displays. *Light: Advanced Manufacturing*, 3(1), 2022. 1

[28] 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*, 2021. 3, 5, 6, 7, 8

[29] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common Objects in 3D: Large-Scale Learning and Evaluation of Real-Life 3D Category Reconstruction. In *Proc. IEEE/CVF Int. Conf. on Comput. Vis.*, October 2021. 2

[30] Tim Salimans, Andrej Karpathy, Xi Chen, and Diederik P. Kingma. PixelCNN++: Improving the PixelCNN with Discretized Logistic Mixture Likelihood and Other Modifications. In *Int. Conf. Mach. Learn.*, 2017. 2

[31] Corentin Sautier, Gilles Puy, Spyros Gidaris, Alexandre Boulch, Andrei Bursuc, and Renaud Marlet. Image-to-Lidar Self-Supervised Distillation for Autonomous Driving Data. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, June 2022. 1

[32] Yichun Shi, Divyansh Aggarwal, and Anil K. Jain. Lifting 2D StyleGAN for 3D-Aware Face Generation. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021. 1

[33] Maxim Tatarchenko, Stephan R. Richter, Rene Ranftl, Zhuwen Li, Vladlen Koltun, and Thomas Brox. What Do Single-View 3D Reconstruction Networks Learn? In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019. 2

[34] Yi Tay, Mostafa Dehghani, Samira Abnar, Yikang Shen, Dara Bahri, Philip Pham, Jinfeng Rao, Liu Yang, Sebastian Ruder, and Donald Metzler. Long Range Arena : A Benchmark for Efficient Transformers . In *Int. Conf. Learn. Representations*, 2021. 4, 5

[35] Michael Tschannen, Eirikur Agustsson, and Mario Lucic. Deep Generative Models for Distribution-Preserving Lossy Compression. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Adv. in Neural Inf. Process. Syst.*, volume 31. Curran Associates, Inc., 2018. 4

[36] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. WaveNet: A Generative Model for Raw Audio. In *9th ISCA Speech Synthesis Workshop*, pages 125–125, 2016. 2

[37] Aaron van den Oord, Nal Kalchbrenner, Lasse Espeholt, koray kavukcuoglu, Oriol Vinyals, and Alex Graves. Conditional Image Generation with PixelCNN Decoders. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Adv. in Neural Inf. Process. Syst.*, volume 29. Curran Associates, Inc., 2016. 2

[38] Aäron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel Recurrent Neural Networks. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proc. of The 33rd Int. Conf. on Mach. Learning*, volume 48 of *Proc. of Mach. Learning Research*, pages 1747–1756, New York, New York, USA, 20–22 Jun 2016. PMLR. 2

[39] Aaron van den Oord, Oriol Vinyals, and koray kavukcuoglu. Neural Discrete Representation Learning. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Adv. in Neural Inf. Process. Syst.*, volume 30, 2017. 2, 4, 5

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Adv. in Neural Inf. Process. Syst.*, volume 30. Curran Associates, Inc., 2017. 3, 4

[41] Dan Wang, Xinrui Cui, Xun Chen, Zhengxia Zou, Tianyang Shi, Septimiu Salcudean, Z. Jane Wang, and Rabab Ward. Multi-View 3D Reconstruction With Transformers. In *Proc. IEEE/CVF Int. Conf. on Comput. Vis.*, 2021. 2

[42] Qian Wang, Neelanjan Bhowmik, and Toby P. Breckon. On the Evaluation of Prohibited Item Classification and Detection in Volumetric 3D Computed Tomography Baggage Security Screening Imagery. In *Proc. Int. Joint Conf. on Neural Networks*, pages 1–8. IEEE, July 2020. 1

[43] Qian Wang and Toby P. Breckon. Contraband Materials Detection Within Volumetric 3D Computed Tomography Baggage Security Screening Imagery. In *Proc. Int. Conf. on Mach. Learning Applications*. IEEE, 2021. 1

[44] Qian Wang and Toby P. Breckon. On the Evaluation of Semi-Supervised 2D Segmentation for Volumetric 3D Computed Tomography Baggage Security Screening. In *Proc. Int. Joint Conf. on Neural Networks*, pages 1–8. IEEE, July 2021. 1

[45] Qian Wang, Najla Megherbi, and Toby P. Breckon. A Reference Architecture for Plausible Threat Image Projection (TIP) Within 3D X-ray Computed Tomography Volumes. *Journal of X-Ray Science and Technology*, 28(3), 2020. 1

[46] Ze Wang, Jiang Wang, Zicheng Liu, and Qiang Qiu. Binary Latent Diffusion. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pages 22576–22585, June 2023. 5

[47] Gengshan Yang, Deqing Sun, Varun Jampani, Daniel Vlasic, Forrester Cole, Ce Liu, and Deva Ramanan. ViSER: Video-Specific Surface Embeddings for Articulated 3D Shape Reconstruction. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Adv. in Neural Inf. Process. Syst.*, volume 34. Curran Associates, Inc., 2021. 1

[48] 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 Recognit.*, pages 10619–10628, 2019. 1, 3, 5, 6, 7, 8

[49] Zhu Zhang, Jianxin Ma, Chang Zhou, Rui Men, Zhikang Li, Ming Ding, Jie Tang, Jingren Zhou, and Hongxia Yang. UFC-BERT: Unifying Multi-Modal Controls for Conditional Image Synthesis. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Adv. in Neural Inf. Process. Syst.*, volume 34, pages 27196–27208. Curran Associates, Inc., 2021. 2

[50] Tinghui Zhou, Philipp Krahenbuhl, Mathieu Aubry, Qixing Huang, and Alexei A. Efros. Learning Dense Correspondence via 3D-Guided Cycle Consistency. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2016. 1