CrossModalityDiffusion: Multi-Modal Novel View Synthesis with Unified Intermediate Representation

Alex Berian, Daniel Brignac, JhihYang Wu, Natnael Daba, Abhijit Mahalanobis
University of Arizona
Tucson, AZ
[berian, dbrignac, jhihyangwu, ndaba, amahalan]@arizona.edu

Abstract

Geospatial imaging leverages data from diverse sensing modalities—such as EO, SAR, and LiDAR, ranging from ground-level drones to satellite views. These heterogeneous inputs offer significant opportunities for scene understanding but present challenges in interpreting geometry accurately, particularly in the absence of precise ground truth data. To address this, we propose CrossModalityDiffusion, a modular framework designed to generate images across different modalities and viewpoints without prior knowledge of scene geometry. CrossModalityDiffusion employs modality-specific encoders that take multiple input images and produce geometry-aware feature volumes that encode scene structure relative to their input camera positions. ***The space where the feature volumes are placed acts as a common ground for unifying input modalities.*** *These feature volumes are overlapped and rendered into "feature images" from novel perspectives using volumetric rendering techniques. The rendered feature images are used as conditioning inputs for a modality-specific diffusion model, enabling the synthesis of novel images for the desired output modality. In this paper, we show that* ***jointly training different modules ensures consistent geometric understanding across all modalities*** *within the framework. We validate CrossModalityDiffusion's capabilities on the synthetic ShapeNet cars dataset, demonstrating its effectiveness in generating accurate and consistent novel views across multiple imaging modalities and perspectives.*

1. Introduction

Geospatial imaging has become increasingly abundant, with datasets spanning a wide range of image sensing modalities such as electro-optical (EO) [16], synthetic aperture radar (SAR) [27], and Light detection and ranging (LiDAR) [10]. Each sensing modality uniquely captures critical features of the Earth's landscape, making them indispensable for comprehensive scene analysis. However, understanding the underlying geometry of scenes from sparse images from various modalities remains a challenging problem.

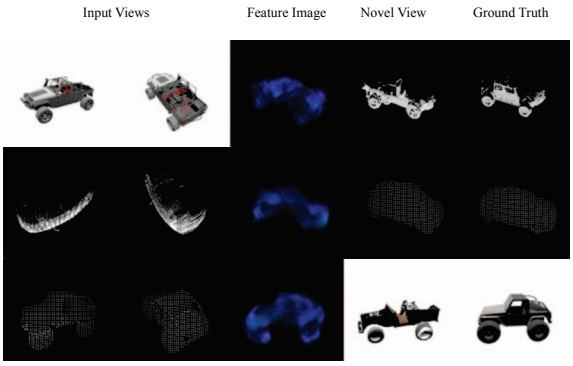

Figure 1. **CrossModalityDiffusion input/output examples.** Three examples of CrossModalityDiffusion with two input images, and an output from a different modality. From the top down: EO to SAR, LiDAR(RA) to LiDAR(P), LiDAR(P) to EO. The corresponding intermediate feature image is shown in the middle, and the ground truth target image is shown on the right.

Neural Radiance Fields (NeRF) [25] revolutionized novel view synthesis (NVS) by learning scene geometry exclusively from images, enabling high-quality view generation. However, NeRF's reliance on dense input images and inability to extrapolate beyond observed viewpoints limits its utility in many scenarios. In contrast, GeNVS [6] explicitly models scene geometry, allowing it to infer structure from sparse inputs. GeNVS uses an encoder to produce geometry-aware feature volumes from each input image, which are aligned and rendered via volume rendering to create a feature image for a novel view. This feature image is then processed by a diffusion model [17, 20] to generate the final output.

While NeRF and GeNVS have shown remarkable results for NVS with EO images, they work on a single imaging modality. Extending NVS to multiple imaging modalities

introduces the problem of multi-modal novel view synthesis (MMNVS), where input and output images may belong to different imaging modalities. Existing approaches like GeNVS face significant limitations in MMNVS due to their dataset-specific nature, requiring retraining for each new modality. Additionally, their tightly coupled architecture restricts generalization.

In this work, we present CrossModalityDiffusion, a modular framework designed for MMNVS. We validate CrossModalityDiffusion on the ShapeNet cars [7] dataset, which we render in EO, perspective LiDAR (LiDAR(P)), range-angle LiDAR (LiDAR(RA)) and SAR modalities. Our results demonstrate that CrossModalityDiffusion effectively synthesizes accurate and consistent novel views across these diverse imaging modalities, showcasing its capability for generalized MMNVS.

In Figure 1 we observe CrossModalityDiffusion operating on EO to SAR, LiDAR(RA) to LiDAR(P), and LiDAR(P) to EO. We see from the feature images in Figure 1 that **no matter the input modality, CrossModalityDiffusion produces geometrically consistent feature volumes.**

CrossModalityDiffusion decouples the three primary components of the GeNVS architecture — (1) the encoder, (2) feature image rendering, and (3) the denoising diffusion model — allowing independent modality-specific modules for input and output modalities. Namely, we use specific encoders and denoisers for each image modality. The feature volumes produced by the encoders are modality-agnostic, allowing for modality fusion. By jointly training the modules within the framework, we ensure the encoder modules learn consistent and transferable geometric representations for the different imaging modalities.

2. Related Work

NeRF [25] revolutionized novel view synthesis by implicitly training a multi-layer perception (MLP) to map 3D coordinates and view directions to color and density for volume rendering. While NeRF achieved state-of-the-art (SOTA) quality in NVS, it requires a dense set of input images to render photo-realistic views. Additionally, NeRF is scene-specific and cannot generalize to unseen targets. Many works incrementally improved NeRF in image quality [2, 3, 14, 41], training time [14, 37], and extended its capabilities [14, 29, 37, 42]. We particularly focus on few-shot NVS, the problem of generating novel views of a scene when given one or a small number of source images. PixelNeRF [44] addressed this problem by training an image encoder with a NeRF module for scene-agnostic performance. GeNVS [6] utilizes a PixelNeRF-like pipeline to produce geometry-aware priors for a powerful diffusion model [20] for few-shot NVS.

Diffusion models are now the standard in high-quality image generation [17, 20, 26, 35, 36] by iteratively denoising the raw image space. Latent diffusion models (LDMs) [21, 31] instead iteratively denoise a latent space representation of data, efficiently generating high-quality images. DreamFusion [28] presented score-distillation-sampling (SDS), which showed that a web-scale text-to-image latent diffusion model [31] implicitly learned 3D information. SDS uses a fixed diffusion model as a critic to train a NeRF. SparseFusion [46] adopts SDS for few-shot NVS by conditioning the LDM with a view-aligned feature grid. Zero-1-to-3 [24, 32] extends the concept in SparseFusion to re-train the LDM with camera pose embeddings to produce better SDS scenes. DreamGaussian [38] uses the concept of SDS for Gaussian splatting [23] scenes instead of a NeRF.

More recently, various works [15, 43] show that when a diffusion model outputs better novel views, those images can be used to directly train a NeRF to get better results than SDS. Multi-view diffusion models, first presented in MVDream [33], utilize cross attention [12, 40] over multiple views of a scene to train a more 3D-aware diffusion model. CAT3D [15] uses a multi-view diffusion model trained at a web-scale to directly produce images for training a NeRF on any scene.

Planar scene representation methods [4, 5, 14, 18, 39], similar to PixelNeRF [44], project render points onto planes to interpolate feature vectors as input to a NeRF model. EG3D [5] trains a Style-GAN [22], to predict 3 planes to represent scenes and fixed NeRF to render the scenes. Large reconstruction models [4, 18, 39] use a similar concept to EG3D but utilize a large visual transformer [12] instead of a GAN.

3. Problem Statement

NVS is the problem of inferring a target image $x_t \in \mathbb{R}^{(128,128,3)}$ from S source images $\{x_{si}\}_{i=1}^{S}$[1]. The target images and source images have associated camera pose matrices $P_t \in \mathbb{R}^{(4,4)}$ and $\{P_{si}\}_{i=1}^{S}$ respectively. We seek to create a NVS model N_θ with parameters θ that predicts novel views

$$\hat{x}_t = N_\theta(P_t, \{x_{si}\}_{i=1}^{S}, \{P_{si}\}_{i=1}^{S}). \quad (1)$$

MMNVS extends the problem of NVS to where the images may come from M different modalities. Each source and target image has an associated modality, (x_{si}, m_{si}) and (x_t, m_t) where $m \in \{1, 2, ..., M\}$. In this paper, we assume modality information is known to the model, so NVS expression in (1) can be rewritten for MMNVS as

$$\hat{x}_t = N_\theta(P_t, m_t, \{x_{si}\}_{i=1}^{S}, \{P_{si}\}_{i=1}^{S}, \{m_{si}\}_{i=1}^{S}). \quad (2)$$

[1]For brevity, we denote $\{x_{si}\}_{i=1}^{S} = \{x_{s1}, x_{s2}, x_{s3}, \ldots, x_{sS}\}$

4. Background: GeNVS

GeNVS [6] encodes each source image x_{si} image into feature volumes $W_i \in \mathbb{R}^{128,128,64,16}$ (point clouds of 16-dimensional features) using a modified DeepLabV3+ [8, 9, 19] segmentation model. The volumes are oriented within the camera field-of-view (frustrum) of their source camera poses P_{si} and a dataset-specific z_{near} to z_{far}.

A 16-channel feature image $F \in \mathbb{R}^{64,64,16}$ of the feature volumes is rendered via volume rendering from a target camera pose P_t. Stratified sampling is used to select points along each ray $\mathbf{r}$. A point $r \in \mathbf{r}$ is trilinearly interpolated in each feature volume to get a latent vector $W_i(r) \in \mathbb{R}^{16}$. The latent vectors for each feature volume are averaged together, then passed through a MLP f to get a 16-channel color c and density σ.

$$c(r), \sigma(r) = f\left(\sum_{i=1}^{S} W_i(r)\right). \tag{3}$$

Once the color and density of every point along a ray $\mathbf{r}$ are predicted, volume rendering is used to predict the 16-channel pixel value $C(\mathbf{r})$. The ray rendering formula follows the NeRF [25] equations

$$C(\mathbf{r}) = \sum_{i=1}^{N} T_i (1 - e^{-\sigma_i \delta_i}) \mathbf{c_i}, \tag{4}$$

where, σ_i and $\mathbf{c}_i$ are the MLP output of the i^{th} sample along the ray from the camera, where

$$T_i = e^{-\sum_{j=1}^{i-1} \sigma_j \delta_j}, \tag{5}$$

and $\delta_i = t_{i+1} - t_i$ is the distance between the i^{th} and $(i+1)^{th}$ samples.

The 16-channel feature image is used as conditioning input to a denoising diffusion [20] U-Net U by concatenating it with 3-channel noise (or noisy target image in training). The denoising U-Net U outputs 3 channels. When sampling new images with multiple denoising steps, the feature image is skip-concatenated to the U-Net's U input. The final output target image $\hat{x}_t$ is the output from the last denoising step.

5. Method

5.1. Multi-Modal Dataset

To better understand how we adapt GeNVS for MMNVS, we must first examine the dataset used to train the framework.

For the model to generalize to unseen images from unseen scenes, we need to first collect images from a variety of scenes. Then from each scene, we need to take a lot of images of it from different viewpoints. From each viewpoint, we need to use different sensors to capture the scene (EO, LiDAR, SAR, etc).

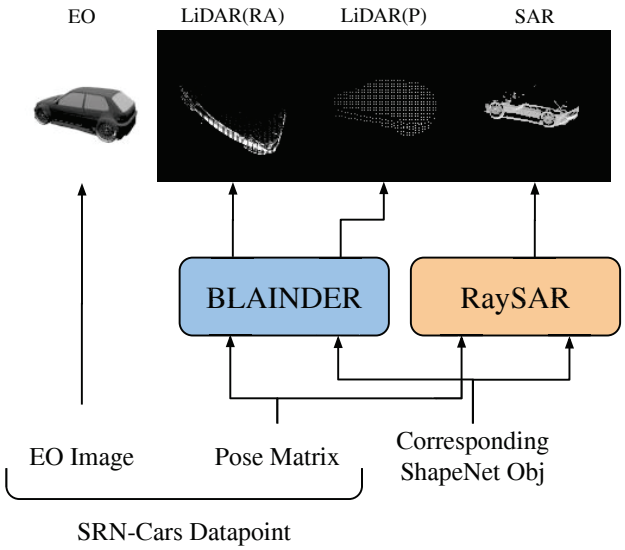

Figure 2. **Dataset Generation.** We begin with the SRN-Cars dataset for EO images and corresponding camera pose matrices. We then use the pose matrix and ShapeNet object file for generating LiDAR(RA) and LiDAR(P) images with BLAINDER, and SAR images with RaySAR

In our experiments, we use the ShapeNet Cars dataset. We get the EO and corresponding poses from the SRN-Cars [34] dataset. We then use BLAINDER [30] to generate LiDAR(RA)/LiDAR(P) images and RaySAR [1] to generate SAR images. This is shown in Figure 2.

5.2. Architecture

GeNVS has three main components. The encoder, feature image generation, and the denoising diffusion model. We begin building the model by training an unmodified GeNVS architecture from scratch exclusively on EO images. This allows for faster learning when training on other modalities. Once the EO-only GeNVS model is pre-trained, we initialize a new encoder and denoiser module for each input and output modality from the pre-trained EO encoder. This architecture is shown in Figure 3.

GeNVS is composed of three major modules: the encoder, the MLP, and the denoiser. By allocating an encoder and denoiser module for each modality, we obtain a modular framework of adapters for MMNVS.

To unify the intermediate representation from the encoders, we only use one MLP to process the overlapping feature volumes created by the different modality encoders. Then according to the target modality, we select the appropriate denoiser for conditional diffusion.

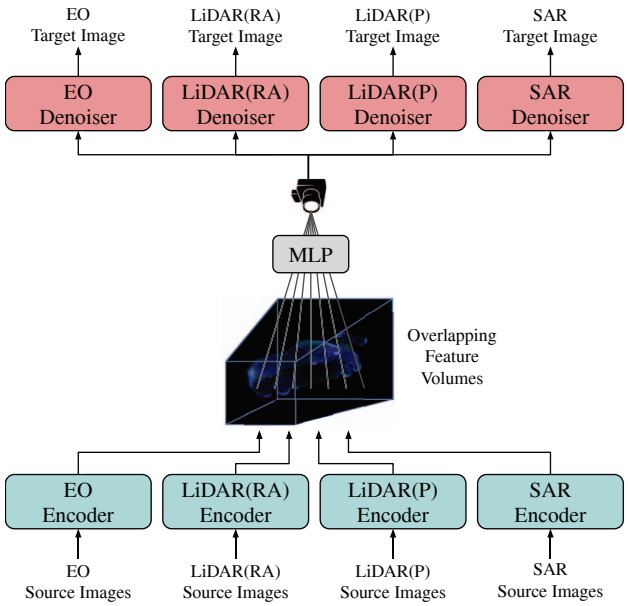

Figure 3. Architecture of CrossModalityDiffusion.

5.3. Joint Training

To allow our framework to input multiple images from different modalities at the same time then output a novel view of the scene in any modality, we jointly train the encoder and denoiser modules for each modality at the same time.

During training, we randomly select between one and three input images from random views and modalities, as well as one random target view and modality. This way, the encoders are incentivized to generate the same feature field despite the different source image modalities. By jointly training the different modules with random input and target modalities, we can create a unified, modality agonistic, implicitly learned intermediate representation of the scene.

Any encoder can encode an image of its modality into the intermediate representation and any denoiser can decode any feature image of a novel view into an image of its modality.

6. Experiments

We follow the distribution of train/validation/test as in the SRN-Cars [34] dataset. EO data is directly taken from the SRN-Cars dataset, and we render the same viewpoints in other modalities. We use RaySAR [1] and BLAINDER [30] for generating SAR and LiDAR images respectively. At present, we use two modes of LiDAR images in the experiments: range-angle (RA), and perspective (P) images. For LiDAR and SAR images, we repeat the single-channel image across three channels for compatibility purposes.

All experiments are conducted on fixed modules trained with the joint training process described in section 5.3. Although the evaluation metrics used are primarily used for EO images, they also function as a useful metric for MM-NVS.

6.1. Single Modality In, Single Modality Out

Results when all input images to CrossModalityDiffusion are of the same modality are shown in Table 1. There is only one output image at a time from the framework, so it is trivial to say all output images are the same modality.

We observe from Table 1 that the easiest of these tasks is EO in EO out (as shown by the 19.66 PSNR). This is because volume rendering captures color information, but not depth. The toughest tasks are any other modality in EO out. This is because the input images from non-EO modalities contain no color information. Images of a black and white car with otherwise consistent geometry would yield a large pixel difference.

In Modality	Out Modality	FID↓	LPIPS↓	DISTS↓	PSNR↑	SSIM↑
EO	EO	19.69	0.13	0.18	19.66	0.87
EO	LiDAR (RA)	29.20	0.19	0.21	17.23	0.77
EO	LiDAR (P)	17.00	0.13	0.17	17.20	0.77
EO	SAR	34.24	0.20	0.20	16.49	0.85
LiDAR (RA)	EO	52.07	0.26	0.27	14.52	0.80
LiDAR (RA)	LiDAR (RA)	28.90	0.18	0.20	17.66	0.78
LiDAR (RA)	LiDAR (P)	16.14	0.12	0.16	17.46	0.78
LiDAR (RA)	SAR	33.42	0.19	0.20	16.64	0.85
LiDAR (P)	EO	44.40	0.25	0.26	14.76	0.80
LiDAR (P)	LiDAR (RA)	24.97	0.15	0.19	19.06	0.81
LiDAR (P)	LiDAR (P)	13.97	0.09	0.15	17.51	0.79
LiDAR (P)	SAR	29.56	0.17	0.19	17.46	0.86
SAR	EO	50.15	0.26	0.26	14.48	0.80
SAR	LiDAR (RA)	27.35	0.18	0.20	17.52	0.78
SAR	LiDAR (P)	16.05	0.11	0.16	17.41	0.78
SAR	SAR	31.81	0.18	0.19	17.17	0.86

Table 1. **Single modality in, single modality out.** We evaluate CrossModalityDiffusion on ShapeNet Cars dataset. To quantify generated image quality, we calculate the Fréchet Inception Distance [13] (FID), Learned Perceptual Image Patch Similarity [45] (LPIPS), Deep Image Structure and Texture Similarity [11] (DISTS), Peak Signal-to-Noise Ratio (PSNR), and Structural Similarity Index Measure (SSIM). ↑ or ↓ indicate better performance from higher or lower values respectively.

6.2. Range-angle Images

Not all sensor modalities produce perspective-projection alike images. We challenge our framework to also work on range-angle images with no modification. We find the denoiser trained on range-angle images still performs well despite no geometry-aware rendering of feature images.

However, we also found it beneficial to render the feature image as a range-angle image. We explicitly calculate the range and angle of all the sampled points and plot it on an empty image. Since the calculated range and angle may

lie between pixels, we use bilinear interpolation weights to scale the feature based on how close it is to each of the four nearest pixels.

6.3. More Images, Better Results

Modality fusion is the process of combining data from different source modalities into the model to provide more reliable, accurate, and useful information. This is a common and important problem in autonomous systems that have multiple sensing modalities such as EO, LiDAR, radar, ultrasound (and so forth) from different locations and orientations. By jointly training the various modality-specific modules of CrossModalityDiffusion, we can fuse information of different modalities from multiple viewpoints into a single, more informative, feature volume representation of the scene.

In Table 2, we evaluate our framework on ShapeNet Cars by randomly selecting S input views from random modalities and comparing the average output image quality (also from random modalities) against when only one of those input images is used. The main conclusion from this experiment is that when more images are used as input, the output image quality improves. This is because CrossModalityDiffusion is fusing its geometric understanding of the scene from each input image. We also observe that as the number of input views increases (lower on Table 2) results improve. Note that the upper number in each row of Table 2 does not change much, because the CrossModalityDiffusion is always using one input image.

Method	FID↓	LPIPS↓	DISTS↓	PSNR↑	SSIM↑
$S = 2$ input views					
Separate	29.99	0.20	0.21	16.40	0.79
Fused	**27.47(-8.4%)**	**0.18(-10.9%)**	**0.20(-5.3%)**	**16.83(+2.7%)**	**0.81(+1.6%)**
$S = 3$ input views					
Separate	28.92	0.20	0.22	16.43	0.79
Fused	**26.03(-10.0%)**	**0.17(-14.7%)**	**0.20(-7.8%)**	**17.15(+4.3%)**	**0.81(+2.2%)**
$S = 4$ input views					
Separate	29.36	0.20	0.21	16.38	0.80
Fused	**25.04(-14.7%)**	**0.16(-16.8%)**	**0.19(-9.0%)**	**17.22(+5.1%)**	**0.82(+2.6%)**
$S = 5$ input views					
Separate	29.23	0.20	0.21	16.52	0.80
Fused	**25.38(-13.2%)**	**0.16(-18.3%)**	**0.19(-9.8%)**	**17.48(+5.8%)**	**0.82(+2.9%)**

Table 2. **Many Vs. One Input image.** In each row the non-bold number for each metric indicates when just one random source image from a random modality is used as input. Note that this number does not change across rows, we simply observe variance. The second number result in bold is when S (indicated by the top left number in the row) random images from random modalities are used as input. For convenience, the number in parenthesis is the difference between the one-input and the S-input results.

6.4. Benefits of Having a Variety of Sensors

Integrating the strengths of various sensing modalities into a unified, more informative representation is highly advantageous in geospatial imaging. As shown in Table 3, leveraging our framework with multiple sensing modalities captured from the same viewpoint enhances performance in the MMNVS task.

Modalities	FID↓	LPIPS↓	DISTS↓	PSNR↑	SSIM↑
LiDAR (RA)	32.343	0.218	0.224	15.911	0.784
SAR	32.204	0.211	0.220	16.042	0.788
LiDAR (P)	41.537	0.183	0.206	16.690	0.805
LiDAR (RA) & SAR	30.697	0.197	0.213	16.332	0.796
LiDAR (RA) & LiDAR (P)	28.536	0.180	0.204	16.749	0.807
SAR & LiDAR (P)	**28.387**	0.181	0.204	16.730	0.807
LiDAR (RA) & SAR & LiDAR (P)	28.542	**0.179**	**0.203**	**16.770**	**0.808**

Table 3. **Different Sensing Modalities at Same Viewpoint.** Results when the input images are all of the same viewpoint with the specified modalities. Random modality outputs are used for this experiment. We observe that more input sensing modalities yields better performance. This is specifically seen in the bottom row of the table.

In this experiment, we evaluate every subset of {LiDAR (RA), SAR, LiDAR (P)} sensing modalities. For each test scene, we randomly select one view as input and three views to predict. The subset of sensing modality images serves as input to CrossModalityDiffusion, which generates novel views for evaluation. We observe the best results in the bottom row of Table 3. As expected, our framework achieves the best performance when all sensing modalities are utilized, even when taken from the same viewpoint. This result underscores the complementary nature of different sensing modalities; each captures unique details of the scene that, when combined, contribute to a richer and more accurate representation.

Another observation we make from Table 3 is that as the number input views increases, the model is using more parameters. In the bottom row of Table 3 more input modalities are used, hence more of the encoder modules as shown in Figure 3 are used. The same cannot be done with just one input modality, because the model would use the same weights multiple times.

6.5. Qualitative Discussion

In Figure 4 we show images from all input/output modalities with two inputs from the same modality. We confirm our hypothesis that EO out from other modalities in is difficult because there is no color information. We see in the EO out examples in Figure 4 that the denoiser outputs a white car when the car is actually yellow. EO to EO does not have this problem.

7. Conclusion

We introduced CrossModalityDiffusion, a modular framework for few-shot MMNVS using a unified intermediate representation. Our experiments demonstrate that CrossModalityDiffusion can effectively transform images across modalities and integrate information from diverse sensing modalities to enrich scene understanding. By

jointly training multiple GeNVS models, we ensure that the learned intermediate representation remains consistent and modality-agnostic. Furthermore, we show that our framework can be adapted to handle non-perspective-projection-based images with minimal modifications.

While CrossModalityDiffusion performs well in MMNVS, its computational demands for training and inference are significant, primarily due to its reliance on a diffusion-based backbone. Despite this limitation, we hope CrossModalityDiffusion inspires new directions in MMNVS and data fusion across sensing modalities. We believe the modality-agnostic intermediate representation has the potential for other downstream tasks beyond MMNVS. Additionally, CrossModalityDiffusion can address scenarios where data availability is imbalanced across sensing modalities, serving as a bridge to overcome such challenges.

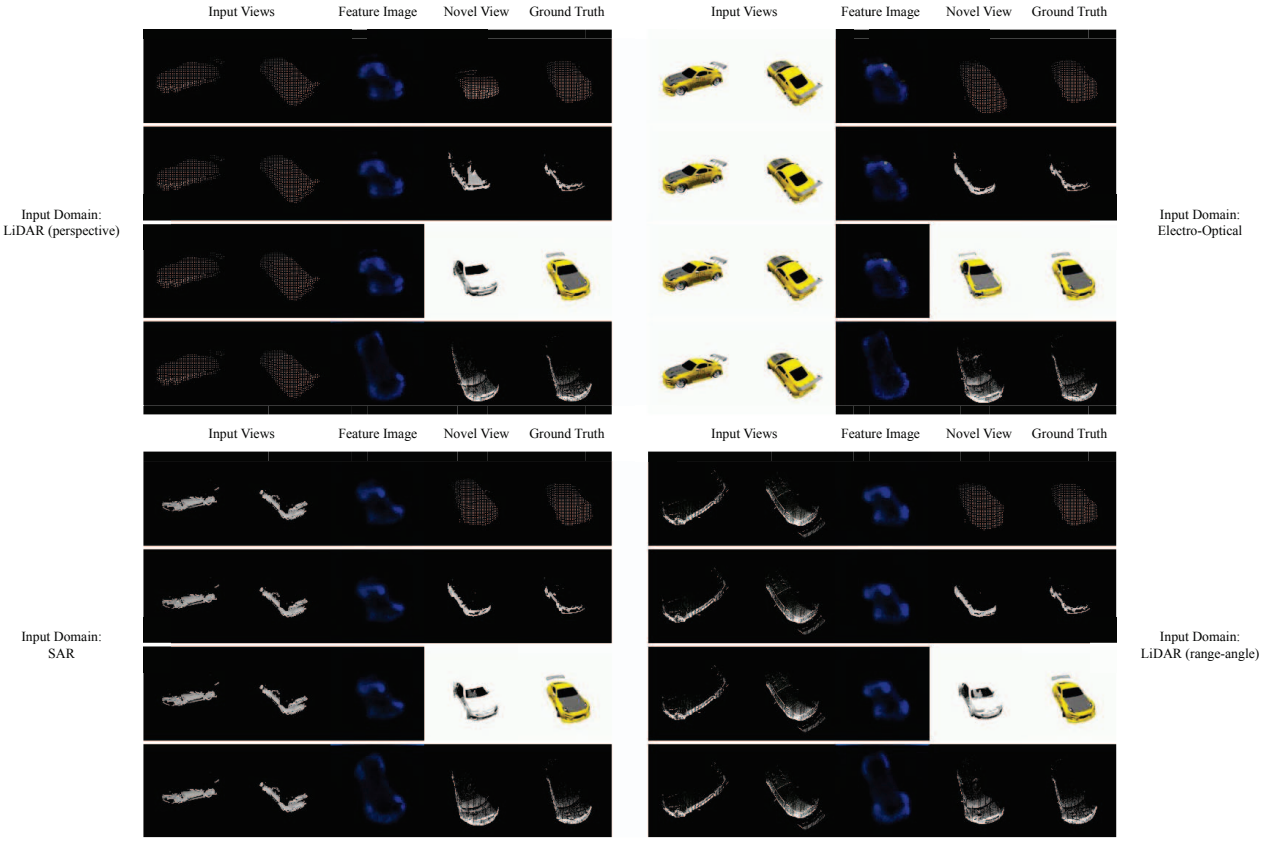

Figure 4. Demonstration of any modality to any modality through unified intermediate representation.

References

[1] Stefan Auer, Richard Bamler, and Peter Reinartz. Raysar-3d sar simulator: Now open source. In *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 6730–6733. IEEE, 2016. 3, 4

[2] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5855–5864, 2021. 2

[3] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Zip-nerf: Anti-aliased grid-based neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19697–19705, 2023. 2

[4] Mark Boss, Zixuan Huang, Aaryaman Vasishta, and Varun Jampani. Sf3d: Stable fast 3d mesh reconstruction with uv-unwrapping and illumination disentanglement. *arXiv preprint*, 2024. 2

[5] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16123–16133, 2022. 2

[6] Eric R Chan, Koki Nagano, Matthew A Chan, Alexander W Bergman, Jeong Joon Park, Axel Levy, Miika Aittala, Shalini De Mello, Tero Karras, and Gordon Wetzstein. Generative novel view synthesis with 3d-aware diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4217–4229, 2023. 1, 2, 3

[7] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 2

[8] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2018. 3

[9] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 3

[10] David Deibe, Margarita Amor, and Ramón Doallo. Big data geospatial processing for massive aerial lidar datasets. *Remote Sensing*, 12(4), 2020. 1

[11] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P. Simoncelli. Image quality assessment: Unifying structure and texture similarity. *CoRR*, abs/2004.07728, 2020. 4

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 2

[13] Maurice Fréchet. Sur la distance de deux lois de probabilité. In *Annales de l'ISUP*, volume 6, pages 183–198, 1957. 4

[14] Sara Fridovich-Keil, Giacomo Meanti, Frederik Rahbæk Warburg, Benjamin Recht, and Angjoo Kanazawa. K-planes: Explicit radiance fields in space, time, and appearance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12479–12488, 2023. 2

[15] Ruiqi Gao, Aleksander Holynski, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul Srinivasan, Jonathan T Barron, and Ben Poole. Cat3d: Create anything in 3d with multi-view diffusion models. *arXiv preprint arXiv:2405.10314*, 2024. 2

[16] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2019. 1

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1, 2

[18] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400*, 2023. 2

[19] Pavel Iakubovskii. Segmentation models pytorch. https://github.com/qubvel/segmentation_models.pytorch, 2019. 3

[20] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022. 1, 2, 3

[21] Tero Karras, Miika Aittala, Jaakko Lehtinen, Janne Hellsten, Timo Aila, and Samuli Laine. Analyzing and improving the training dynamics of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24174–24184, 2024. 2

[22] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 2

[23] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 2

[24] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9298–9309, 2023. 2

[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):99–106, 2021. 1, 2, 3

[26] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR, 2021. 2

[27] Wolfgang Pitz and David Miller. The terrasar-x satellite. *IEEE Transactions on Geoscience and Remote Sensing*, 48(2):615–622, 2010. 1

[28] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv*, 2022. 2

[29] Mohamad Qadri, Michael Kaess, and Ioannis Gkioulekas. Neural implicit surface reconstruction using imaging sonar. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1040–1047. IEEE, 2023. 2

[30] Stefan Reitmann, Lorenzo Neumann, and Bernhard Jung. Blainder—a blender ai add-on for generation of semantically labeled depth-sensing data. *Sensors*, 21(6):2144, 2021. 3, 4

[31] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2

[32] Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view diffusion base model, 2023. 2

[33] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023. 2

[34] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. *Advances in Neural Information Processing Systems*, 32, 2019. 3, 4

[35] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 2

[36] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 2

[37] Matthew Tancik, Ben Mildenhall, Terrance Wang, Divi Schmidt, Pratul P Srinivasan, Jonathan T Barron, and Ren Ng. Learned initializations for optimizing coordinate-based neural representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2846–2855, 2021. 2

[38] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *arXiv preprint arXiv:2309.16653*, 2023. 2

[39] Dmitry Tochilkin, David Pankratz, Zexiang Liu, Zixuan Huang, Adam Letts, Yangguang Li, Ding Liang, Christian Laforte, Varun Jampani, and Yan-Pei Cao. Triposr: Fast 3d object reconstruction from a single image. *arXiv preprint arXiv:2403.02151*, 2024. 2

[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, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 2

[41] Dor Verbin, Peter Hedman, Ben Mildenhall, Todd Zickler, Jonathan T. Barron, and Pratul P. Srinivasan. Ref-NeRF: Structured view-dependent appearance for neural radiance fields. *CVPR*, 2022. 2

[42] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *NeurIPS*, 2021. 2

[43] Rundi Wu, Ben Mildenhall, Philipp Henzler, Keunhong Park, Ruiqi Gao, Daniel Watson, Pratul P Srinivasan, Dor Verbin, Jonathan T Barron, Ben Poole, et al. Reconfusion: 3d reconstruction with diffusion priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21551–21561, 2024. 2

[44] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4578–4587, 2021. 2

[45] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 4

[46] Zhizhuo Zhou and Shubham Tulsiani. Sparsefusion: Distilling view-conditioned diffusion for 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12588–12597, 2023. 2