Consistent Multimodal Generation via A Unified GAN Framework

Zhen Zhu
UIUC
zhenzhu4@illinois.edu

Yijun Li
Adobe Inc.
yijli@adobe.com

Weijie Lyu
UIUC
wlyu3@illinois.edu

Krishna Kumar Singh
Adobe Inc.
krishsin@adobe.com

Zhixin Shu
Adobe Inc.
zshu@adobe.com

Sören Pirk
Adobe Inc.
soeren.pirk@gmail.com
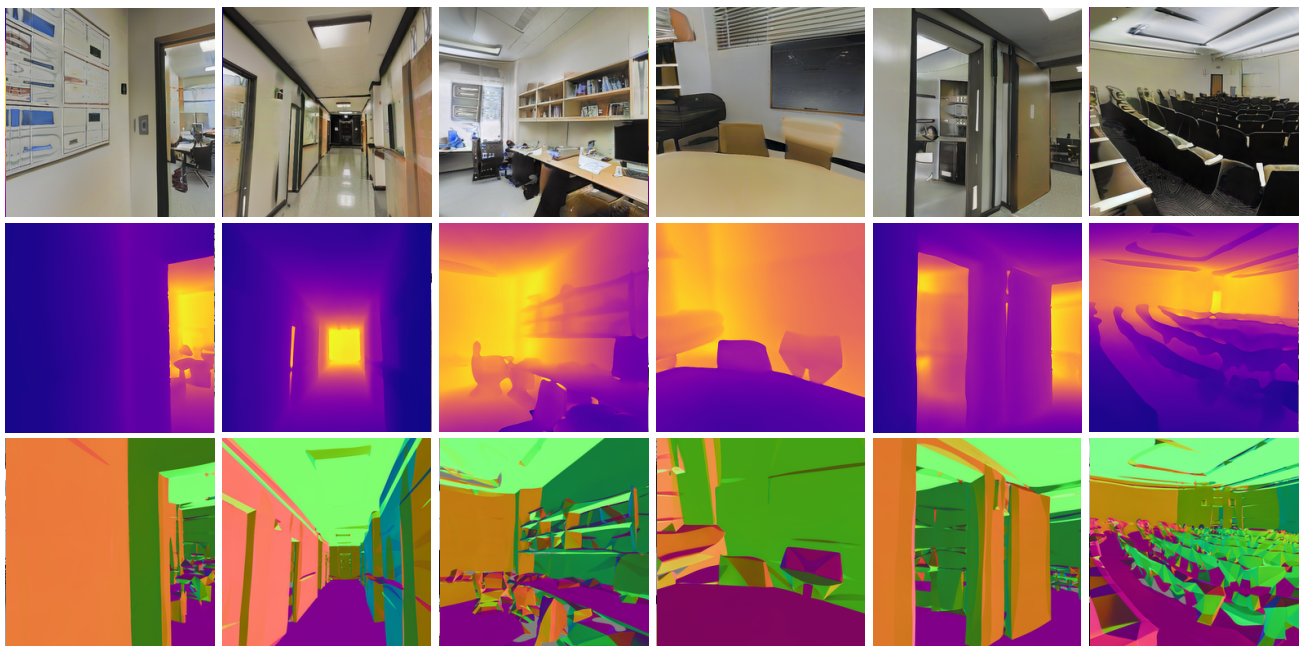
Derek Hoiem
UIUC
dhoiem@illinois.edu

Figure 1. Generated samples of our MultimodalGAN. Results from top to bottom row are RGB, depth and surface normals respectively. Each column is generated from the same latent code in one run and shows consistency between each modality. These samples cover diverse classes in Stanford2D3D dataset [1] and exhibit multiple camera poses even without using camera parameters.

Abstract

We investigate how to generate multimodal image outputs, such as RGB, depth, and surface normals, with a single generative model. The challenge is to produce outputs that are realistic, and also consistent with each other. Our solution builds on the StyleGAN3 architecture, with a shared backbone and modality-specific branches in the last layers of the synthesis network, and we propose per-modality fidelity discriminators and a cross-modality consistency discriminator. In experiments on the Stanford2D3D dataset, we demonstrate realistic and consistent generation of RGB, depth, and normal images. We also show a training recipe to easily extend our pretrained model on a new domain, even with a few pairwise data. We further evaluate the use of synthetically generated RGB and depth pairs for training or fine-tuning depth estimators. Code will be available at here.

1. Introduction

Realistic synthetic data is useful for training, especially when supervision is difficult to collect, such as for scene geometry. Ability to generate RGB images has progressed tremendously [10, 14–16], but, without additional scene information, synthetic RGB images are of limited use for training. Our goal is to produce images in multiple modali-

ties that are realistic and mutually consistent. Our key idea is that, just as various scene properties (e.g. RGB values and depth) stem from an underlying scene and viewpoint, we can generate them with a single model based on a shared underlying representation. By sharing representations among related tasks, we enable our model to generalize better on the target task.

We build on the StyleGAN3 [14] architecture. A single latent code is used to generate a shared backbone representation, which is refined by branches to generate each modality. Different from previous works that use certain modality-related losses, we employ only adversarial losses, which offers flexibility to scale up to many modalities without considering modality specifics. This approach eliminates the need for laborious loss weight tuning associated with modality-related losses. We introduce two types of discriminators, the fidelity and consistency discriminators, to encourage the realism and consistency among the output modalities. Furthermore, instead of specifying different data augmentation policies for different modalities, we adopt a single data augmentation strategy across all modalities, ensuring our framework's scalability and simplicity.

We demonstrate the use of our **MultimodalGAN** by applying it to a set of challenging tasks. First, we show that multiple modalities can realistically be generated together and with a high degree of consistency. Second, we apply our model in a multi-modality adaptation setting, where we generate modalities for a new domain. Considering that a source domain contains abundant pairs of RGB and depth images, while the target domain has abundant RGB images but very few or no depth data, our model can be used for generating non-RGB data in the target domain. Third, we use our framework to synthesize depth estimation dataset that comprises of RGB images and the corresponding depth maps, which can be used to train depth estimation models. Evaluation results on the challenging Stanford2D3D [1] real dataset show that our framework is capable of synthesizing more realistic multi-modality outputs compared to state-of-the-art approaches.

To summarize, our main contributions include: 1) We devise an unconditional GAN framework that generates multiple modalities simultaneously. We showcase that this MultimodalGAN supports smooth and consistent transitions by traversing the latent space. 2) We propose a training recipe that extends the MultimodalGAN pretrained on a domain to another domain even with only a few training pairs. 3) We demonstrate our model can serve as a powerful data engine to synthesize RGB images and corresponding depth labels to improve the performance on depth estimation.

2. Related Work

Generative image modeling. Generative modeling by learning the data distribution has initiated a considerable amount of interests that has led to remarkable progress in recent years. Seminal approaches range from variational autoencoder [18] and generative adversarial networks [3, 10] to flow-based models and denoising diffusion probabilistic models [7]. Due to larger datasets and increasingly more powerful compute-availability, the quality of the results these methods generate steadily increases. However, most of these generative models are learning the distribution from the RGB modality, only to synthesize realistic images – other data modalities are rarely considered.

Here we highlight a few approaches that also address other modalities in addition to generating RGB images. The work of S^2-GAN [26] proposes a two-stage sequential framework by first generating surface normals and then the RGB result for indoor scenes. In contrast, DatasetGAN [32] focuses learning semantic segmentation maps along with RGB images in a few-shot generation scenario. After learning the GAN model on the RGB modality, it is required to label a few segmentation maps for the generated images to then fine-tune an additional network to produce segmentation labels. SemanticGAN [19] also tackles this task by generating RGB and segmentation masks concurrently. It relies on an additional segmentation model to provide supervision for mask generation whereas in our case we utilize the generic adversarial loss that is agnostic to domains and do not rely on pre-trained models. Noguchi *et al*. [21] introduce RGBD-GAN, which is based on unsupervised 3D representation learning from 2D images, without using depth data but by relying on a 3D consistency loss. Polymorphic-GAN [17] learns morph maps on shared features, enabling style transfers only in the RGB domain. However, we focus on generating dense image analysis such as depth and surface normal. To make the generation more 3D-aware, a recent work [23] employs two generators to synthesize RGB and depth. However, instead of using real depth data, they leverage a pretrained depth estimator [28] to extract synthetic depth as the training data, which is consequently limited by the depth range learned in the estimator. Unlike the existing approaches, we do not focus on a particular modality, but address the simultaneous generation of multiple modalities by only using one single network.

Multi-task learning. It has been recognized that the joint training on multiple data modalities enables learning more powerful representations [4,5,20,30,35]. Classic multi-task learning [6, 33] aims at simulating the learning process of human beings based on the human intelligence of integrating knowledge across domains. The general idea of multi-task learning is to train machine learning models with data from a set of related tasks simultaneously, in order to discover a shared representations. Oftentimes, configuring and training a multi-task architecture is challenging as hyperparameters and losses need to be carefully tuned jointly. The seminal approach introduced by Zamir *et al*. [31] proposes

a fully computational approach for modeling the structure of the space of visual tasks to quantify the relationship between them. One recent work [2] proposes to use generative approaches for visual multi-task learning, by coupling a discriminative multi-task network with a generative network. However, it aims to predict multi-task outputs from RGB images and leverages the generative network to synthesize diverse images to facilitate training. Unlike previous works focusing on discriminative tasks, our model generates the RGB image along with its pixel-level dense representation (e.g., depth) directly from randomly sampled latent codes.

3. Method

3.1. Overview

As shown in Fig. 2, starting from a random latent $z \in \mathcal{R}^l$ of length l drawn from standard Gaussian distribution $\mathcal{N}(0, 1)$, the generator G aims to convert z to generate different modalities:

$$[g_m] = G(z), \ \ m \in \{\text{RGB}, \text{depth}, \text{normal}\}, \quad (1)$$

where $[\cdot]$ indicates concatenation of outputs in the channel dimension, g_m is the generation of a particular modality represented by m. This derives the unconditional multi-modality generator. In order to produce realistic outputs for multiple modalities, the generator should have adequate capacity. One of the most powerful unconditional generators in the literature is StyleGANs [14–16] which shows smooth latent interpolation and high generation quality. The structure of our generator is innovated by StyleGAN3 [14].

A StyleGAN generator is comprised of a mapping network and a synthesis network. In StyleGAN3, the depth of the mapping network is reduced from 8 to 2. We follow the same design as it introduces less parameters and reduces the effort of optimization. The mapping network transforms the input latent z to an intermediate latent space $w \in \mathcal{W}$. w is passed to the affine transformation layers A of all synthesis blocks in the synthesis network, adjusting the styles accomplished inside each block. We maintain most of the designs for the synthesis network in StyleGAN3, including Fourier features to replace the constant input, discarding noise input to ensure transformation hierarchy, normalizing features by their exponential moving average estimates, *etc*. More details can be found in the supplementary material.

3.2. Generator

The diagram of our generator is depicted in Fig. 2. We keep the same depth as StyleGAN3. The commonly shared part of the generator enumerates until layer 8. Then we design separate branches for the respective modalities. This design considers that different modalities share the same underlying structure of a real scene and we assume that the
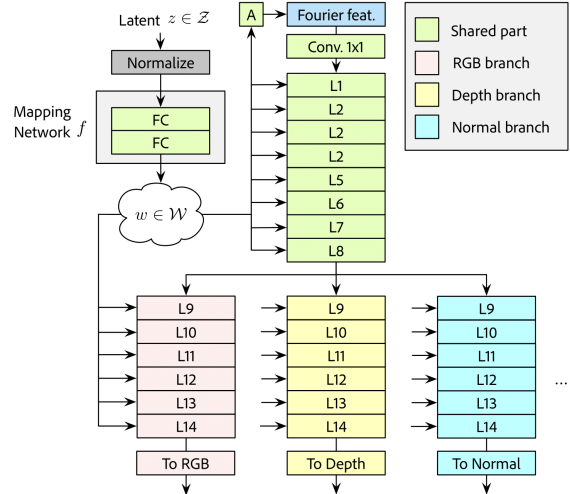

Figure 2. Diagram of our generator, which employs a shared backbone first and separate branches to generate each modality.

shared part of our architecture is able to represent the similarities, even without clear guidance. The different branches share the same structure, except for the output layer that uses different numbers of channels for each of the modalities. The outputs of all modalities are grouped as tuples.

3.3. Discriminators

We aim to generate data that is realistic and consistent across different modalities. Ideally, this goal should be achieved through only one discriminator that receives the concatenated modalities. However, the generator can trivialize the learning of this discriminator by feeding spatially aligned tuples, while some modalities may not look realistic at all. This problem becomes more severe when dealing with more modalities (see Sec. 4). Therefore, it is necessary to use multiple discriminators that validate the fidelity of each modality separately. However, only using these standalone discriminators also cannot guarantee consistency of the generated tuples.

Therefore, we developed two types of discriminators. The first type is called *Fidelity Discriminators* (denoted by D_F^m) used for assessing the fidelity of the corresponding generated modality m. The other type of discriminator is called *Consistency Discriminator* (denoted by D_C). This discriminator judges whether a tuple of inputs is consistent or not. The input to D_C is the concatenation of all modalities. A diagram of our discriminators is shown in Fig. 3. The architecture of each discriminator is generally the same except for different input channels of the first convolutional layers. The structure is inherited from the discriminator design in StyleGAN2 [16] and later versions [13, 14].
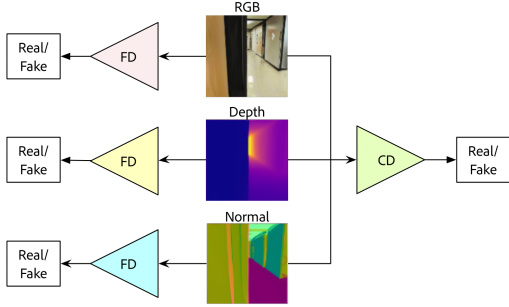
Figure 3. Diagram of our discriminators. "CD" represents consistency discriminator and "FD" represents fidelity discriminator of a specific modality.

3.4. Training

Leveraging supervised training for unconditional generators is usually difficult for a single domain. It becomes feasible in multi-task learning as one can use a synthesized sample of one modality to estimate other modalities, while ensuring consistency. For example, if the generator is trained to produce RGB and depth images simultaneously, we can employ a pre-trained depth estimator [28] to predict depth images from the synthesized RGB images and supervise the synthesized depths images to be closer to estimated ones. However, this approach relies on the performances of the respective models and it may fail when the pre-trained models are trained on a different data distribution. Furthermore, relying on complex exiting architectures increases the compute requirements, which prevents scaling to multiple data modalities. Therefore, we mainly use adversarial losses to train the whole system for simplicity and reliability. During training, all discriminators are trained jointly and their outputs are concatenated together:

$$D_o(x) = [[D_F^m(x^m)], \ D_C(x)], \quad (2)$$

where x^m is either generated or real data of a modality m and x is the concatenation of either a generated or real tuple, which means that $x = [x^m]$. By default, we use the non-saturating logistic loss [10]. Hence, the full objective is:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x \sim p_{\text{data}}(\mathbf{x})}[\log D_o(\mathbf{x})] \\ - \mathbb{E}_{\mathbf{z} \sim p_{\mathbf{z}}(\mathbf{z})}[\log D_o(G(\mathbf{z}))]. \quad (3)$$

3.5. Data Augmentation Strategy

Collecting data for different tasks is commonly a laborious effort. When performing multimodal generation training, pairwise data is even harder to obtain, which poses great challenges for training our model. In order to cope with data insufficiency, a commonly used technique is data augmentation. As shown in other methods [13, 34], applying data augmentation operations to real and generated data

is beneficial to ensure a balance between discriminator and generator, even when the amount of data is small. We also use data augmentation to train our model, but we face the challenge that it is not possible to use the same augmentation operations across the different data modalities. For example, when generating RGB images and normals jointly, performing a horizontal flip for both modalities does not produce a reliable tuple – images of surface normals cannot be simply flipped.

However, as shown in [13], as long as the discriminator is trained on an adequate number of samples of real data (no augmentations are performed) it can help the generator to approach the real data distribution. Therefore, we apply the same augmentations for all modalities, except for color distortions of RGB images. Even with this seemingly unreasonable augmentation strategy, our results show the consistency of our generated modalities. We apply the same augmentation pipeline as in [13], since it contains more geometric translations compared to other approaches. We find that these translations are important for modalities capturing geometry and structure. Each augmentation operation in the pipeline has the same execution probability p of being executed or $1 - p$ of not being executed. We follow the same approach as described in their paper to adjust p. Moreover, we treat p as a signal to stop training when it reaches a threshold, i.e., 0.7 in our case, since larger values of p also indicates severe overfitting of the discriminator. When p exceeds the threshold, it also becomes increasingly rare for the generator to see true training samples, thus causing generation with distortions.

3.6. Cross-modal Fine-tuning

In certain domains, there are often abundant RGB images yet scarce corresponding labels (e.g., depth). For instance, a construction site may possess ample RGBD pairs for most rooms (domain A), but insufficient pairs for a specific scene, such as a utility room (domain B). This data imbalance makes training a generator from scratch challenging. To address this, we employ cross-domain fine-tuning, wherein a model pretrained on domain A is fine-tuned to accommodate domain B, despite limited pairwise data.

In our framework, only the consistency discriminator necessitates pairwise supervision, and therefore, is exclusively trained on pairwise data. Conversely, all available data can be used to train the fidelity discriminators. However, the limited pairwise data may skew the training of different discriminators, empirically leading to inconsistent generation or even optimization divergence. To rectify this imbalance, we modify the data sampling procedure to sample paired and unpaired data equally. The paired data provides accurate annotations that assist the generator in generalization, while the unpaired data introduces diversity to the corresponding modalities. Through the comprehensive use of all

data, we enable successful generalization to a novel domain.

4. Experiments

4.1. Dataset

We focus on the Stanford2D3D [1] dataset for evaluation. It contains 70,496 pairwise inputs of multiple data modalities, such as RGB, depth, surface normal, etc. Images in this dataset contain different types of complex scenes with various objects (e.g., chair, shelves, monitor) inside a Stanford building, rather than a single object located at the center of an image. Previous works [23] mainly experimented with LSUN-Bedroom and LSUN-Kitchen [29] and show that networks based on StyleGAN can generate realistic images on these datasets, while ours is the first to show the capacity of StyleGAN generators to deal with such a complex dataset. Please refer to the supplementary material for implementation details.

4.2. Evaluation Metrics

Since our model is unconditional, it is not feasible to directly compare generated and real samples in a pairwise manner. Instead we adopt FID [11] to calculate the distance between distributions of generated data and real data. We feed data of a particular modality to an InceptionV3 [24] model pre-trained on ImageNet [22] to obtain feature means and variances. Then, we calculate the Frechet distance using these statistics averaged over a fraction of sampled data points. For real data, we sample $\min(50K, |\mathcal{D}_{\text{train}}|)$ real data points, where $|\mathcal{D}_{\text{train}}|$ indicates the number of data points in the training dataset. For calculating the statistics in generated data, we generate 50K data points.

Besides assessing the distribution discrepancy, we also want to ensure that the different modalities are consistent with each other. In order to evaluate consistency, we first generate 5K pairwise samples. For depth, we calculate the averaged Scale Invariant Depth Error (**SIE**) [9] between a generated depth map and an estimated depth map, obtained by feeding the generated RGB to a pre-trained depth estimator [28]. For surface normals, we use an off-the-shelf surface normal estimator [8, 12] to estimate surface normals from a generated RGB image and then compare the result with our generated normals. To assess normal consistency with generated RGB images, we use the standard metrics proposed in prior work [25]: the mean and median of angular error measured in degrees, and the percentage of pixels whose angular error is within γ degrees.

4.3. Comparison

We compare with methods that address unconditional synthesis of multiple modalities. To this end, the Depth-GAN [23] and RGBD-GAN [21] are most closely related to our work. We use their original codebases to run experiments on Stanford2D3D dataset and keep their original model and settings for a fair comparison. DatasetGAN [32] is based on a StyleGAN generator for learning additional tasks, such as segmentation and key-point estimation. The underlying assumption is that the StyleGAN generator already provides a geometrical or disentangled understanding of the generated image. Therefore, the only effort is to extract labels from intermediate representations of the generator; a few images need to be generated and then manually annotated. Finally, they trained a small network on top of StyleGAN generator to translate features into dense labels. However, the efficacy of DatasetGAN is only demonstrated for a single-object dataset. A limitation of this approach is that some modalities cannot be annotated by human annotators – for example, depth is captured instead of annotated. Therefore, we cannot directly compare with DatasetGAN. To validate our model, we first train a depth estimation model [28] using all real RGB and depth pairs. Then we pre-train a standard StyleGAN3 [14] model on RGB images. After that, we build a depth branch on top of the pre-trained generator, which results in an identical architecture to our generator with RGB and depth branches. We fix all parameters except for those in the depth branch and leverage supervised training to train the depth branch by treating the estimated depth from the LeReS model on generated RGB image as ground-truth. We denote this approach as "RGB→DEPTH". LeReS [28] also provides well-trained models on lots of data. We also use these pre-trained representations to provide guidance for training the depth branch and compare the performance with our approach (RGB→DEPTH*).

In Tab. 1 we show that our method achieves good RGB-FID and the best Depth-FID and SIE scores, demonstrating the effectiveness of our approach. Both the RGB→DEPTH model and RGB→DEPTH* model have very large Depth-FIDs and SIEs. The reason lies in the incompatibility between the inputs to the depth branch and depth modality: either the feature maps from L8 or the intermediate latent codes w are optimized towards the RGB modality, posing large challenge when learned towards depth, especially other parts are fixed except for the depth branch (See Fig. 2). This further confirms the benefit of learning multiple modalities jointly. In Fig. 4, both RGBD-GAN and DepthGAN fail to generate plausible depth, while RGBD-GAN is comparatively better at RGB images. RGB→DEPTH and RGB→DEPTH* produce better depth results, but are still far from satisfaction compared to ours. This consolidates the effectiveness of our approach in generating multiple modalities with one model.

4.4. Ablation

To further explore the efficacy of our approach we provide a set of ablation studies. Most of the ablation experi-
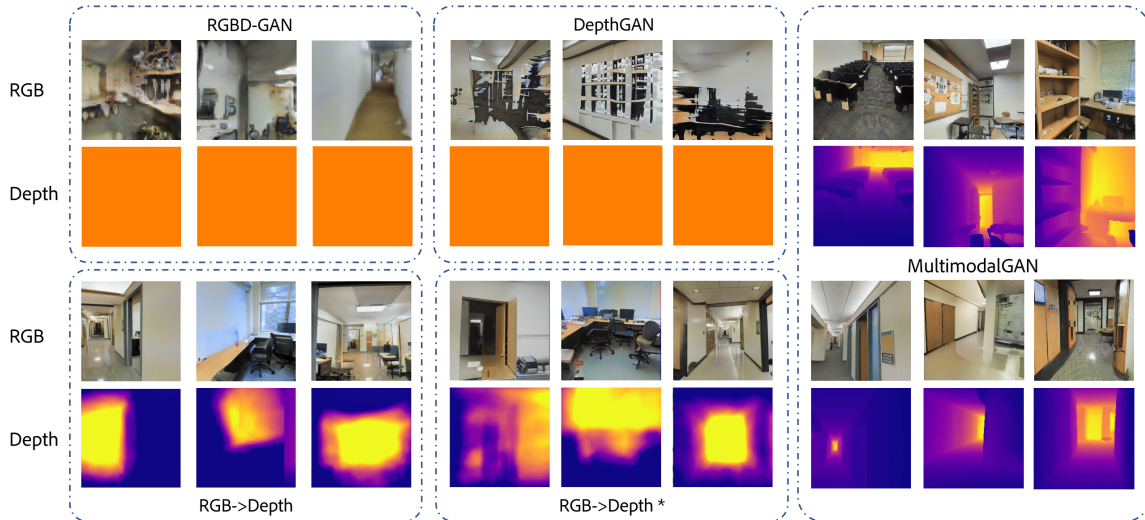
Figure 4. Qualitative comparison between the proposed MultimodalGAN and previous methods. All generated depth maps are visualized using the same color map.

Table 1. Quantitative comparisons with other methods. For all metrics, the lower, the better.

Method	RGB-FID	Depth-FID	SIE
RGBD-GAN [21]	105.0	46.7	0.339
DepthGAN [23]	326.3	182.9	0.467
RGB→DEPTH	**13.8**	227.5	1.745
RGB→DEPTH*	**13.8**	245.9	1.182
MultimodalGAN	14.6	**24.8**	**0.192**

Table 2. Ablation in terms of data augmentation strategies.

Model	Aug.	RGB-FID	Depth-FID	SIE
RGB	No aug.	41.6	-	-
RGB	Diff. [34]	29.2	-	-
RGB	ADA [13]	12.6	-	-
Full	No aug.	35.3	33.3	0.236
Full	Diff. [34]	20.6	29.9	0.193
Full	ADA [13]	14.6	24.8	0.192

ments are conducted on the RGB and depth modalities.

Choice of augmentation strategies. As discussed in Sec. 3.5, data augmentation is vital, especially when the amount of data is not sufficient for effectively training a modoel. Multiple discriminator augmentation strategies can be used to improve performance of GAN training, including DiffAugment [34] and ADA [13]. For a comparison of different augmentation methods, we use two models: 1) RGB baseline: we directly train a StyleGAN3 model on RGB images from the Stanford2D3D dataset; 2) Full model: our final model as shown in Fig. 2 and Fig. 3. For the full model, we apply the same augmentation strategy to both modalities. As shown in Tab. 2, ADA provides the best results for both models. Our hypothesis is that the performance improvement is mainly caused by geometric translations in ADA. In our case, the depth modality requires

Table 3. Ablation in terms of generator architecture on Stanford2D3D dataset.

Method	RGB-FID	Depth-FID	SIE
RGB baseline	12.6	-	-
Depth baseline	-	20.9	-
RGBD baseline	15.5	26.2	0.189
Branch baseline	13.6	22.9	0.192

modelling scene structures, which can be enhanced by geometric translation operations. As shown in [13], geometric translations are helpful when the amount of data is small.

Building separate branches for different modalities. Our generator contains a shared part and then branches out for the respective data modalities. Alternatively, one can build only one branch with the final output layer to produce multi-channel outputs, which can be split to represent different modalities. We refer to this baseline as **RGBD baseline** when producing RGB and depth simultaneously. This baseline can be trained with one discriminator which receives a four-channel input (first three channels for RGB, the rest one for depth). As another baseline, we also train a StyleGAN3 model on the RGB and depth domains, which we refer to as **RGB baseline** and **Depth baseline**. These models are all trained with one discriminator. For fair comparison, we build a branch generator with only one discriminator, called **Branch baseline**, for which we build separate branches from L9 as shown in Fig. 2.

The quantitative comparison is shown in Tab. 3. For the multi-modality baselines, both the RGB-FIDs and Depth-FIDs are lower than those of the single-modality baselines. However, the RGB baseline and the Depth baseline alone do not have many practical applications compared to the multimodal generation models. When comparing the RGBD baseline with the Branch baseline, we observe better results

Table 4. Ablation in terms of the layer indexes to build branches.

Layer index	RGB-FID	Depth-FID	SIE
L6	13.3	23.5	0.207
L9	14.6	24.8	0.192
L12	15.3	25.6	0.191

for the Branch baseline in both FID metrics (with a slightly inferior SIE). This shows the advantage of generating data with better quality and decent consistency.

Layer indexes to build branches. After the generator architecture, we investigate the correct place to build branches in the generator. The RGBD baseline can be treated as an extreme case when regarding the branching index as L13. So we further investigate smaller indices. We experimented with defining branching from L6, L9 and L12 along with the different spatial sizes of their output features.

As illustrated in Tab. 4, if the starting branch is L6, both FID measurements are comparatively inferior to others, since this model fails to converge. Our hypothesis is that the formation of shared underlying structures is not complete in the shallow layers. As training starts, the outputs of different modalities do not align well, causing the consistency discriminator to easily differentiate real/fake pairs which consequently leads to unsuccessful training. Since the generation quality of RGB is not satisfactory, the SIE metric is high. Using L9 already provides reasonable results and we observe that the qualitative results from L12 show more higher frequency artifacts (the depth modality contains high frequency noise). Therefore, we choose L9 to build branches as default.

Discriminator design. In our full model, we use the fidelity discriminators (FD) and the consistency discriminator (CD) to ensure both generation quality and modality consistency. We aim to show the effect of each type to the performance by removing either type from the Full model. In Tab. 5, we show that when the model only has two modalities, using only the consistency discriminator would suffice. However, when we scale up to more modalities (adding surface normals), using both types of discriminators provide the best performance. Using only the fidelity discriminators cannot guarantee a consistent tuple across modalities, as indicated by the large SIE.

Scale up to more modalities. Our model is not constrained to only two modalities. As shown in Tab. 5, we can further scale up to also generate surface normals. Training with three modalities can be accomplished without a significant drop in performance for the modalities, when using both consistency discriminator and fidelity discriminators. As we can see from the result in Fig. 1, the results generated by this model show diversity over many scene classes of Standford2D3D dataset and complex layouts.
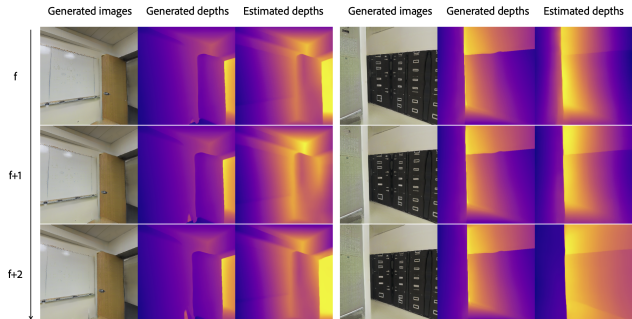

Figure 5. Qualitative results showing the smoothness of generated depths versus estimated depths from [28]. "f" represents the the initial frame index.

5. Applications

5.1. Consistent RGBD Generation

One advantage of StyleGAN-based approaches is their capacity to generate smooth outputs through linear interpolation in the latent space. Our model retains this capability across not only the conventional RGB modality but also additional modalities. This is an obvious advantage of our approach over existing methods that estimate corresponding modalities from each RGB image, as these estimations can exhibit considerable flickering issue in subsequent interpolations. As demonstrated in Fig. 5, our generated depths maintain consistent layouts and local details across all time frames, whereas the estimated depths show substantial inconsistencies in extensive regions in successive frames. More video results can be found in the supplementary materials.

5.2. Cross-domain Fine-tuning

To demonstrate cross-domain fine-tuning, we partition the Stanford2D3D dataset into two subsets, designating the "auditorium" scene (comprising 6,371 pairwise data) as the target domain, and the remaining scene classes as the source domain. In the target domain, we employ only a percentage $\mathbf{p}\%$ of the pairwise data, assuming the model has access to the remaining $(1 - \mathbf{p}\%)$ RGB images. The source domain undergoes pre-training to concurrently generate RGB images and depth. Upon completion, we apply the fine-tuning procedure outlined in Sec. 3.6 to the pre-trained model on the target domain. We select $\mathbf{p}$ from the set $5, 10, 15$ to exhibit the model's performance with varying amounts of pairwise data for fine-tuning. As illustrated in Tab. 6, all metrics decrease with the incorporation of more pairwise data. According to the RGB-FID, 10% of RGBD pairs suffice for fine-tuning. We also present qualitative results and corresponding videos in the supplementary material, showing the model's satisfactory generalization to a new scene, exhibiting considerable latent smoothness and quality, even with minimal fine-tuning.

Table 5. Ablation in terms of discriminator architecture. *w.* normal represents whether to add the normal modality into the generation task.

		RGB	Depth		Surface Normal					
						Anglular Error° ↓		% Within γ° ↑		
Disc.	*w.* normal	RGB-FID	Depth-FID	SIE	Normal-FID	Mean	Median	11.25°	22.5°	30°
FD		14.1	25.3	0.298	-	-	-	-	-	-
CD		13.6	22.9	0.192	-	-	-	-	-	-
CD+FD		14.6	24.8	0.192	-	-	-	-	-	-
CD	✓	185.7	48.5	0.360	44.0	20.36	17.23	32.24	62.73	77.39
CD+FD	✓	15.6	29.0	0.195	15.4	14.77	12.23	46.40	79.69	89.87

Table 6. Fine-tuning on held-out auditorium scene using different portions of pairwise data. The first two columns represent the portion of images used for training.

RGB+D	RGB	RGB-FID	Depth-FID	SIE
5%	95%	66.3	115.4	0.164
10%	90%	53.8	102.3	0.158
15%	85%	53.0	91.3	0.144

5.3. Generating Datasets for Depth Estimation

Can we generate a synthetic dataset to benefit downstream tasks? To explore this objective, we synthesize a dataset of RGB and depth images with our multi-modal GAN and validate its usage in the task of depth estimation.

We first split the Stanford2D3D dataset into training and validation sets using a ratio of 6:1. Of the training samples, 20% (12,045 RGB+D pairs) are randomly chosen to train our generative model. Despite the small training set, the baseline achieves good performance: the RGB-FID, Depth-FID and SIE of this model are 23.7, 27.0 and 0.192, respectively. Then, we use the model to produce a series of synthetic datasets, of one to five times the initial 20% set. We train LeReS [28] from scratch on each of these datasets for 30K iterations, since the models converge well within this amount of training. We measure performance with absolute mean relative error (AbsRel) and Weighted Human Disagreement Rate (WHDR) [27] for evaluation. More details are provided in the supplementary material. As shown in Fig. 6, using more of our generated data consistently improves both metrics. When the total number of generated and real samples is equal to the size of the full training set ($r + 4g$ on the x-axis), the model achieves similar performance to using 80% of the real data. This is impressive, given that the generator is only trained with 20% of real data, and indicates that our generative model can effectively generate synthetic training data.

6. Conclusion and Limitations

We introduce a novel architecture to generate multiple image modalities simultaneously. We demonstrate its effectivenss on the challenging Stanford2D3D dataset and several interesting applications. Especially, our architecture can predict the modalities coherently, enabling synthesis
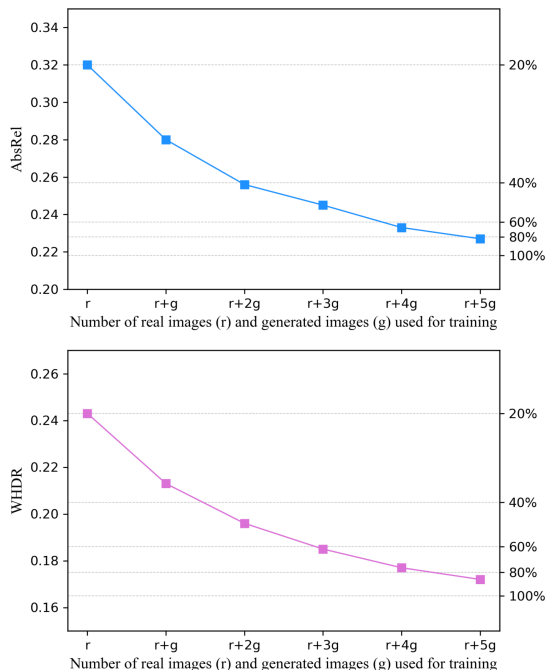

Figure 6. Depth estimation performance of LeReS [28] when trained on different combinations of real data and generated data. The x-axis represents the combination of real and generated images. The left y-axes are the numbers of AbsRel or WHDR obtained by the corresponding models, respectively. The dashed horizontal lines indicate the performances of training on different portions of training dataset. "r"/ "g" represent the numbers of real/generated images. In this case, $r = g = 12,045$, which is 20% size of the training set.

of datasets that, in turn, can be employed to enhance the performance of downstream tasks. Currently, we attempted generating RGB, depth and surface normal under the GAN framework. In the future, we would like to try more modalities such as segmentation mask and other generative models like diffusion models. For generality, we do not consider leveraging modality-specific characteristics, which might be chances for further improvements, such as 3D constraints for more geometrically consistent pairs.

Acknowledgement

References

[1] Iro Armeni, Sasha Sax, Amir R Zamir, and Silvio Savarese. Joint 2d-3d-semantic data for indoor scene understanding. *arXiv preprint arXiv:1702.01105*, 2017. 1, 2, 5

[2] Zhipeng Bao, Martial Hebert, and Yu-Xiong Wang. Generative modeling for multi-task visual learning. In *Proc. ICML*, 2022. 3

[3] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 2

[4] Arunkumar Byravan and Dieter Fox. Se3-nets: Learning rigid body motion using deep neural networks. In *ICRA*, 2017. 2

[5] Vincent Casser, Sören Pirk, Reza Mahjourian, and Anelia Angelova. Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos. In *AAAI*, 2019. 2

[6] Michael Crawshaw. Multi-task learning with deep neural networks: A survey. *arXiv preprint arXiv:2009.09796*, 2020. 2

[7] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, 2021. 2

[8] Ainaz Eftekhar, Alexander Sax, Jitendra Malik, and Amir Zamir. Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. In *CVPR*, 2021. 5

[9] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *NeurIPS*, 2014. 5

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 1, 2, 4

[11] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017. 5

[12] Oğuzhan Fatih Kar, Teresa Yeo, Andrei Atanov, and Amir Zamir. 3d common corruptions and data augmentation. In *CVPR*, 2022. 5

[13] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In *NeurIPS*, 2020. 3, 4, 6

[14] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. In *NeurIPS*, 2021. 1, 2, 3, 5

[15] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 1, 3

[16] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, 2020. 1, 3

[17] Seung Wook Kim, Karsten Kreis, Daiqing Li, Antonio Torralba, and Sanja Fidler. Polymorphic-gan: Generating aligned samples across multiple domains with learned morph maps. In *CVPR*, pages 10620–10630, 2022. 2

[18] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2

[19] Daiqing Li, Junlin Yang, Karsten Kreis, Antonio Torralba, and Sanja Fidler. Semantic segmentation with generative models: Semi-supervised learning and strong out-of-domain generalization. In *CVPR*, pages 8300–8311, 2021. 2

[20] Yao Lu, Sören Pirk, Jan Dlabal, Anthony Brohan, Ankita Pasad, Zhao Chen, Vincent Casser, Anelia Angelova, and Ariel Gordon. Taskology: Utilizing task relations at scale. In *CVPR*, 2021. 2

[21] Atsuhiro Noguchi and Tatsuya Harada. Rgbd-gan: Unsupervised 3d representation learning from natural image datasets via rgbd image synthesis. *arXiv preprint arXiv:1909.12573*, 2019. 2, 5, 6

[22] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015. 5

[23] Zifan Shi, Yujun Shen, Jiapeng Zhu, Dit-Yan Yeung, and Qifeng Chen. 3d-aware indoor scene synthesis with depth priors. *arXiv preprint arXiv:2202.08553*, 2022. 2, 5, 6

[24] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016. 5

[25] Xiaolong Wang, David F. Fouhey, and Abhinav Gupta. Designing deep networks for surface normal estimation. In *CVPR*, 2015. 5

[26] Xiaolong Wang and Abhinav Gupta. Generative image modeling using style and structure adversarial networks. In *ECCV*, 2016. 2

[27] Ke Xian, Chunhua Shen, Zhiguo Cao, Hao Lu, Yang Xiao, Ruibo Li, and Zhenbo Luo. Monocular relative depth perception with web stereo data supervision. In *CVPR*, 2018. 8

[28] Wei Yin, Jianming Zhang, Oliver Wang, Simon Niklaus, Long Mai, Simon Chen, and Chunhua Shen. Learning to recover 3d scene shape from a single image. In *CVPR*, 2021. 2, 4, 5, 7, 8

[29] Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, and Jianxiong Xiao. LSUN: construction of a large-scale image dataset using deep learning with humans in the loop. *CoRR*, abs/1506.03365, 2015. 5

[30] Amir R. Zamir, Alexander Sax, Nikhil Cheerla, Rohan Suri, Zhangjie Cao, Jitendra Malik, and Leonidas J. Guibas. Robust learning through cross-task consistency. In *CVPR*, 2020. 2

[31] Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *CVPR*, 2018. 2

[32] Yuxuan Zhang, Huan Ling, Jun Gao, Kangxue Yin, Jean-Francois Lafleche, Adela Barriuso, Antonio Torralba, and Sanja Fidler. Datasetgan: Efficient labeled data factory with minimal human effort. In *CVPR*, 2021. 2, 5

[33] Yu Zhang and Qiang Yang. A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, 2021. 2

[34] Shengyu Zhao, Zhijian Liu, Ji Lin, Jun-Yan Zhu, and Song Han. Differentiable augmentation for data-efficient GAN training. In *NeurIPS*, 2020. 4, 6

[35] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G. Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, 2017. 2