# Semantic RGB-D Image Synthesis

Shijie Li
University of Bonn
Bonn, Germany
lishijie@iai.uni-bonn.de

Rong Li
HKUST (Guangzhou)
Guangzhou, China
rongli@hkust-gz.edu.cn

Juergen Gall
University of Bonn
Bonn, Germany
gall@iai.uni-bonn.de

## Abstract

*Collecting diverse sets of training images for RGB-D semantic image segmentation is not always possible. In particular, when robots need to operate in privacy-sensitive areas like homes, the collection is often limited to a small set of locations. As a consequence, the annotated images lack diversity in appearance and approaches for RGB-D semantic image segmentation tend to overfit the training data. In this paper, we thus introduce semantic RGB-D image synthesis to address this problem. It requires synthesising a realistic-looking RGB-D image for a given semantic label map. Current approaches, however, are uni-modal and cannot cope with multi-modal data. Indeed, we show that extending uni-modal approaches to multi-modal data does not perform well. In this paper, we therefore propose a generator for multi-modal data that separates modal-independent information of the semantic layout from the modal-dependent information that is needed to generate an RGB and a depth image, respectively. Furthermore, we propose a discriminator that ensures semantic consistency between the label maps and the generated images and perceptual similarity between the real and generated images. Our comprehensive experiments demonstrate that the proposed method outperforms previous uni-modal methods by a large margin and that the accuracy of an approach for RGB-D semantic segmentation can be significantly improved by mixing real and generated images during training.*

## 1. Introduction

RGB-D semantic segmentation is essential for mobile agents as it enables gaining a precise perception of the environment. While there are fast methods like ESANet [25] that are suitable for robotics applications, the collection of annotated RGB-D training data remains a bottleneck. This is in particular an issue for robots in homes since the data collection is restricted due to privacy concerns, but it is also an issue for UAVs or delivery robots that need to enter private property.

(a) Real (b) RGB-D OASIS (c) SCMIS (Ours)

Figure 1. Comparison of images that are generated for a given label mask. The left column shows the real RGB-D image that corresponds to the label mask. The middle column shows RGB-D images generated by OASIS extended to RGB-D. The proposed approach (right column) generates more realistic RGB-D images than OASIS. While the depth maps of OASIS are inaccurate and noisy and the edges of the tables are not straight, our approach generates images where the RGB and depth image are consistent.

An interesting direction to increase the diversity of training data is thus to generate training data from label maps as shown in Fig. 4. This task is also known as semantic image synthesis [18, 24, 28] where a semantic label map is provided and the aim is to generate realistic images that are consistent with the label map. While this also requires training data, it allows the generation of each annotated training image variant where the appearance of the present objects, walls, and floor differs from the original image as shown in Fig. 4. While previous works focused on uni-modal semantic image synthesis, i.e., only generating RGB images, we propose a multi-modal approach that generates realistic-looking RGB-D images from semantic label maps. We thus call the task semantic RGB-D image synthesis. In contrast to previous works, we not only evaluate the quality of the generated images but also demonstrate that we can significantly improve the accuracy of an RGB-D semantic segmentation approach [25] by mixing real and generated images during training as shown in Fig. 6.

Although approaches for uni-modal semantic image synthesis like OASIS [24] can also be applied to RGB-D im-

ages by treating RGB and depth as a single modality, they do not perform well as shown in Fig. 1. We therefore propose an approach that explicitly considers RGB and depth as two different modalities. However, since appearance and geometry are highly correlated, we aim to disentangle the modal-independent information that is encoded in the semantic label map and the modal-dependent information that is needed to generate images for both modalities, namely depth and color.

To this end, the generator uses an encoder that is modal-independent and two modal-dependent decoders for both modalities as illustrated in Fig. 2. The decoding is done gradually from the down-sampled label map with spatially-adaptive normalization, which is conditioned on the modal-independent information for each scale. To train the model, we ensure that the generated depth maps correspond to the real depth maps. Furthermore, we propose a discriminator that ensures semantic consistency between the generated image and the semantic label map. We call the approach thus **S**emantic **C**onsistent **M**ulti-Modal **I**mage **S**ynthesis (SCMIS).

Since we also aim that the RGB images to look as realistic as possible, we enforce that the features of generated and real images are similar. While this can be theoretically achieved by the perceptual loss [13], we show that the perceptual loss impedes learning and has a negative impact on the results. This is due to the commonly used VGG features that do not necessarily separate the semantic classes. We, therefore, integrate the perceptual loss into the discriminator such that the VGG features are adapted and separate the classes better.

We evaluate our approach for multi-modal semantic image synthesis on standard RGB-D datasets where it outperforms uni-modal approaches by a large margin. Furthermore, we show that the generation of depth also improves the quality of the generated RGB images. The contributions of this work can be summarized as: 1) We propose a novel semantic consistent multi-modal generator that generates realistic RGB-D images by disentangling the modal-independent and modal-dependent information. 2) We propose an adaptive alignment discriminator which performs well for uni-modal (RGB) and multi-modal (RGB-D) semantic image synthesis. 3) We conduct comprehensive experiments and the results demonstrate that the proposed method outperforms previous methods by a large margin. 4) We demonstrate that the accuracy of an RGB-D semantic segmentation approach can be improved by mixing the real training images with the generated images.

## 2. Related Work

**Generative Adversarial Networks.** GAN [7] are commonly used for semantic image synthesis and many improvements have been proposed over the years. One main issue is the instability in adversarial training. [20] improved the training stability by designing a robust architecture and providing some design guidelines. Different from [20], some methods [1, 15, 23] made the training more stable by using some tricks like adding regularizers or proposing better loss functions. In addition, one can easily improve the efficiency of networks by existing model compression techniques [35, 36].

**Semantic Image Synthesis.** Pix2Pix [11] was one of the first approaches for semantic image synthesis and used a general image-to-image translation framework with a conditional GAN architecture [16]. Since Pix2Pix [11] can only handle images of the small resolution, Pix2PixHD [33] proposed a multi-resolution approach to generate high-resolution images. SPADE [18] removed the common normalization layers like instance normalization [31] and instead modulates the image content spatially by a learned affine transformation. Inspired by SPADE [18], many semantic image synthesis methods have been proposed [12, 24, 29, 30]. Among these methods, OASIS [24] showed that a segmentation-based discriminator provides much more precise supervision and improves the results substantially.

**Geometric Image Synthesis.** Recently, some works explore 3D data generation given 2D supervision. Most of them focus on specific shapes like faces [5, 6, 19]. Some works go further [10, 17] and reconstruct some simple 3D models from a single image. Apart from reconstructing 3D models, some methods generate geometric-consistent images. For instance, [39] generates a bird view from a frontal view in the context of a driving car. [4, 34] utilize geometric information to improve the quality of generated images. While these works generate geometry or utilize geometry for generating images, none of them generates multi-modal data. We will also show that uni-modal approaches for semantic image synthesis do not perform well for semantic RGB-D image synthesis.

## 3. Semantic Consistent Multi-Modal Image Synthesis

While previous works for semantic image synthesis generate only one modality, namely an RGB image, for a given label map, we address multi-modal semantic image synthesis. Instead of generating only an RGB image from a given label map, our approach generates an RGB-D image where color and depth are treated as two modalities as shown in Fig. 2. An important aspect of the architecture is that modal-independent information of the semantic layout and modal-dependent information that is needed to generate an RGB and a depth image are separated.

### 3.1. Semantic Consistent Multi-Modal Generator

Previous methods for semantic image synthesis considered only one modality and naively extending these meth-

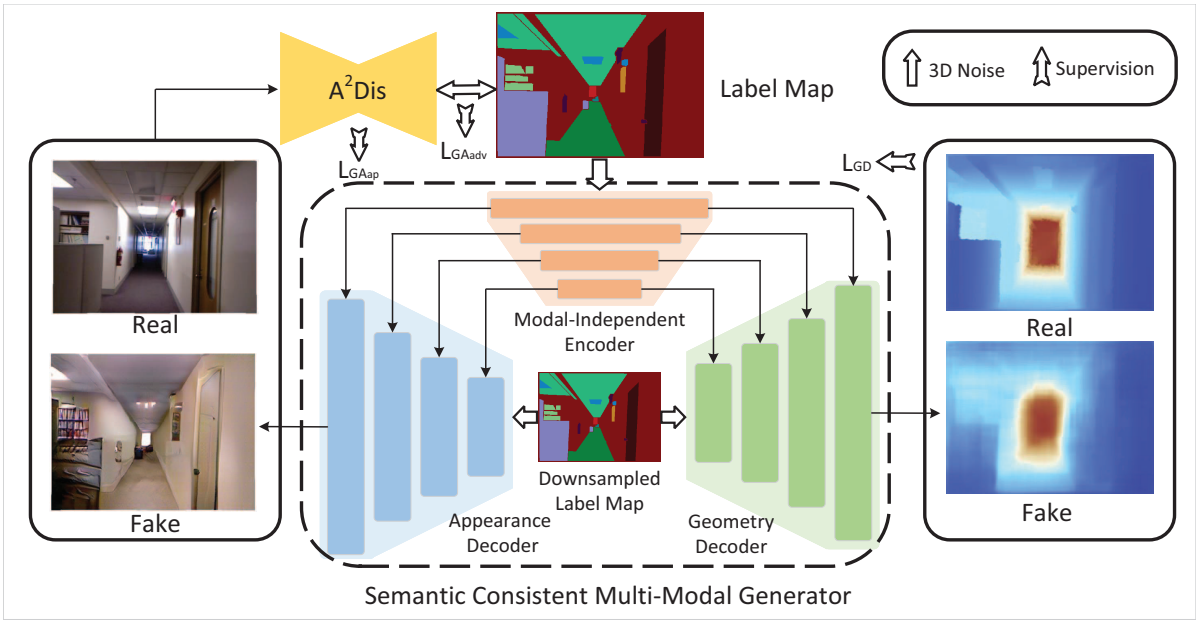
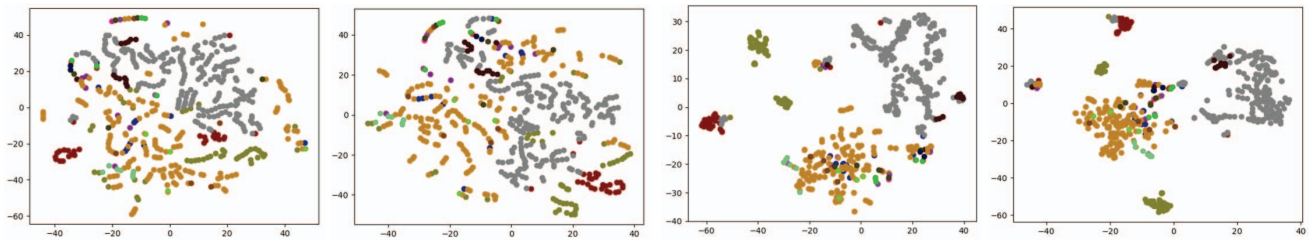



Figure 2. Overview of the proposed **S**emantic **C**onsistent **M**ulti-Modal **I**mage **S**ynthesis (SCMIS) approach. The generator takes a label map as input and generates an RGB image (blue) and depth image (green) that are semantically consistent with the input label map. For training, we use for each modality a different loss term. While $\mathcal{L}_{GD}$ measures the quality of the generated depth map, the proposed Adaptive Alignment Discriminator ($A^2$Dis) measures the quality of the generated RGB image and consistency with the input label map.



(a) VGG-19 pretrained on ImageNet [14] (b) Adaptative VGG-19

Figure 3. (a) Visualization of the features that are used for the perceptual loss, i.e., VGG-19 pre-trained on ImageNet [14]. The features are projected by t-SNE [32]. The left plot shows the features for real images and the right plot for generated images. The colors correspond to different semantic classes. (b) Visualization of the features that are learned for adaptive perceptual loss.

ods to multi-modal data does not result in good results as shown in Fig. 1. Indeed, we will demonstrate in the experiments that generating RGB-D images from semantic label maps in the same way as generating RGB images is even slightly worse than generating both modalities RGB and depth independently. However, depth and color are highly correlated and should not be treated independently. We therefore address the research question of how RGB-D images can be better generated from semantic label maps.

Our proposed multi-modal generator shown in Fig. 2 takes into account that both modalities share the same scene properties like layout or object locations, which are independent of the modality. The modal-independent encoder (red) thus encodes the common properties that are shared by both modalities. The decoders, however, are modal-dependent and we use one appearance decoder (blue) and one geometry decoder (green), which generate the RGB image and the depth image, respectively. We now describe the encoder and both decoders more in detail.

As shown in Fig. 2, the modal-independent encoder takes as input a channel-wise concatenation of a semantic label map and a 3D noise tensor. The noise tensor allows to generate various plausible images from the same label map. The encoder consists of blocks with a convolution layer, a batch normalization layer, and a ReLU. The resolution decreases after each block.

The modal-dependent decoders have the same architecture and generate from a downsampled version of the semantic map concatenated with the 3D noise tensor the RGB image or depth map, respectively, at the resolution of the original input label map. Each decoder consists of ResNet blocks [8] with spatially-adaptive normalization (SPADE)

[18]:

$$\gamma^i_{x,y,c}(\mathbf{e}^i)\frac{h^i_{x,y,c,n}-\mu^i_c}{\sigma^i_c}+\beta^i_{x,y,c}(\mathbf{e}^i) \tag{1}$$

where $n$ is the batch size, $c$ is the channel, $(x, y)$ is a pixel at layer $i$. $h^i_{x,y,c,n}$ is the input to the normalization layer, and $\mu^i_c$ and $\sigma^i_c$ are the mean and standard deviation of $h^i_{x,y,c,n}$ for channel $c$, i.e., computed over $n$, $x$, and $y$. The spatially-adaptive normalization can be conditioned on a tensor $\mathbf{e}$.

We use the tensor $\mathbf{e}$ to inject the modal-independent information at each scale for the decoders. $\mathbf{e}^i$ is thus the tensor of the encoder at layer $i$. In this way, both decoders share the same semantic and modal-independent information and gradually generate from the downsampled input map the two different modalities.

### 3.2. Adaptive Alignment Discriminator

Before we describe the full loss to train the generator, we first discuss the second contribution, which is denoted by Adaptive Alignment Discriminator ($A^2$Dis) in Fig. 2. The discriminator serves two purposes. It aims to ensure the semantic consistency of the generated image with the label map and aims to ensure that the generated images look realistic.

To this end, we use a discriminator that predicts per-pixel probabilities of $N + 1$ classes which correspond to $N$ semantic classes and one additional 'fake' class as in [24]. For the discriminator, we will evaluate different types in our experiments as shown in Fig. 7. The network is trained by using the ground-truth label map for a real image and labelling all pixels as 'fake' for a generated image. To deal with imbalanced classes, a weighted $N+1$-class cross-entropy loss is used:

$$\begin{aligned}\mathcal{L}_{GA_{adv}} = &-\mathbf{E}_{(\mathbf{x},\mathbf{l})}\left[\sum_{c=1}^{N}\alpha_c\sum_{x,y}^{H\times W}\mathbf{l}_{x,y,c}\log D(\mathbf{x})_{x,y,c}\right]\\ &-\mathbf{E}_{(\mathbf{z},\mathbf{l})}\left[\sum_{x,y}^{H\times W}\log D(G(\mathbf{z},\mathbf{l}))_{x,y,c=N+1}\right]\end{aligned} \tag{2}$$

where $\mathbf{l}$ is the semantic label map, $\mathbf{x}$ is a real image, and $(\mathbf{z}, \mathbf{l})$ is the concatenation of the 3D noise tensor and the semantic label map. The weight $\alpha_c$ is computed as the inverse of the per-pixel class frequency to give rare classes higher weights:

$$\alpha_c = E_{\mathbf{l}}\left[\frac{H\times W}{\sum_{x,y}^{H\times W}\mathbf{l}_{x,y,c}}\right]. \tag{3}$$

By this design, the semantic consistency among generated images are enforced.

Although this discriminator can train a good generator, it does not ensure that the generator produces images that follow the real data distribution as no constraint between generated images and real images exists. To solve this issue, we can complement our discriminator with the perceptual loss [13], which is widely used for training generative networks [18]. But this will introduce two issues. First, an additional pre-trained VGG-19 model is required which is inefficient. Second, the used VGG-19 model is pre-trained on another dataset and might not be suitable for our task. This domain gap leads to semantic inseparable features as can be seen in Fig. 3 (a). As consequence, the perceptual loss works against the loss (2) since it pushes the generated features towards a feature space where the classes are not separable and difficult to classify.

The proposed adaptive alignment discriminator solves these issues by unifying the semantic consistency among generated images and the alignment between real and generated images in the same framework. It is designed as a segmentation-based discriminator with a learnable VGG-19 as the backbone. In this way, the learned features fit better to the task and separate the semantic classes better as shown in Fig. 3 (b). To maintain feature alignment between real and generated images, the perceptual loss is applied to the learned features, and thus named adaptive perceptual loss:

$$\mathcal{L}_{GA_{ap}} = \frac{1}{N}\sum_{i=1}^{N}||r_i - f_i||_1 \tag{4}$$

where $i$ denotes the layer number, $N$ is the total number of layers, and $r_i$ and $f_i$ denote the features from the real and generated image, respectively.

As mentioned at the beginning of this section, we investigated three architectures for the discriminator with the adaptive VGG-19 backbone (AVGG). As shown in Fig. 7, it includes a simple architecture with upsampling, pyramid pooling [38], and a U-Net architecture [22]. As we will show in the experiments, the architecture with pyramid pooling performs best. Hence, the total loss for the appearance decoder is given by:

$$\mathcal{L}_{GA} = \mathcal{L}_{GA_{adv}} + \mathcal{L}_{GA_{ap}}. \tag{5}$$

To recover the geometric structure, we apply the L1 loss on the output of the geometry decoder:

$$\mathcal{L}_{GD} = \frac{1}{N}\sum_{i=1}^{N}||d_i - \hat{d}_i||_1 \tag{6}$$

where $N$ is the number of pixels, $\hat{d}_i$ is the generated depth value, and $d_i$ is the ground truth depth value for pixel $i$. In addition, we also include the LabelMix regularization [24]:

$$\begin{aligned}\mathcal{L}_{LM} = &||D_{logits}(LabelMix(\mathbf{x},\hat{\mathbf{x}},M))\\ &- LabelMix(D_{logits}(\mathbf{x}), D_{logits}(\hat{\mathbf{x}}), M)||_2^2\end{aligned} \tag{7}$$

where $D_{logits}$ are the logits before the last softmax and $M$ is a binary mask for mixing real and generated images $(\mathbf{x}, \hat{\mathbf{x}})$

| Methods | NYU | | SUNRGBD | |
|---|---|---|---|---|
| | FID ↓ | mIoU$_D$ ↑ | FID ↓ | mIoU$_D$ ↑ |
| SPADE [18] | 122.1 | 47.2 | 246.6 | 27.8 |
| TSIT [12] | 125.4 | 51.8 | 130.8 | 40.5 |
| DAGAN [29] | 99.2 | 51.4 | 57.4 | 51.2 |
| OASIS [24] | 77.5 | 58.3 | 26.4 | **62.2** |
| RGB-D OASIS | 89.1 | 54.7 | 25.1 | 57.8 |
| SCMIS | **55.2** | **59.8** | **19.9** | 60.7 |

Table 1. Comparison on multi-modal image synthesis.

by

$$LabelMix(\mathbf{x}; \hat{\mathbf{x}}; M) = M \odot \mathbf{x} + (1 - M) \odot \hat{\mathbf{x}}. \quad (8)$$

The entire loss is thus given by

$$\mathcal{L} = \mathcal{L}_{GA} + \mathcal{L}_{GD} + \mathcal{L}_{LM}. \quad (9)$$

# 4. Experiment

We evaluate our method on the two challenging datasets: SUN RGB-D [27] and NYU [26]. The NYU [26] dataset includes 1449 RGB-D images with 40 classes while the SUN RGB-D [27] dataset contains 10355 RGB-D images with 37 classes. We also evaluate our approach for uni-modal semantic image synthesis on the Cityscapes [2] dataset which includes 2975 training images and 500 validation images with 35 classes. All experiments have been performed on a single TITAN RTX with a fixed random seed. The source code will be released upon acceptance.

We will first evaluate the quality of the generated images in Section 4.1 and compare our multi-modal approach to uni-modal approaches for semantic image synthesis. In Section 4.2, we will demonstrate that our approach can be used to augment the training data of an approach for RGB-D semantic segmentation by mixing real and generated images. Finally, we evaluate the impact of the loss terms and architecture design in Section 4.3.

## 4.1. Semantic RGB-D Image Synthesis

We use the common Fréchet Inception Distance (FID) [9] to evaluate the quality of the generated images using the implementation from [24]. Lower FID means better image quality. To measure the semantic alignment between the generated image and the semantic label map (semantic consistency), we use mean Intersection-over-Union (mIoU). To this end, we use the RGB-D semantic segmentation framework ESANet [25] with pre-trained weights. For the evaluation, the generated images with real-depth images are used as input. For uni-modal semantic image synthesis, we use DRN [37] to maintain consistency with previous works. As we use two networks for computing mean Intersection-over-Union, 'mIoU$_D$' in the tables denotes that ESANet [25] for RGB-D images has been used and 'mIoU' denotes that DRN [37] for RGB images has been used.

We compare our method to other methods for uni-modal semantic image synthesis on two RGB-D datasets. The results are shown in Table 4.1. We observe that our method produces more realistic images and achieves a lower FID score. Adapting uni-modal semantic image synthesis methods to semantic RGB-D image synthesis leads to worse performance (RGB-D OASIS). This means that these architectures cannot handle multi-modal information well. The qualitative results in Fig. 4 also show that our method generates more realistic images compared to previous methods.

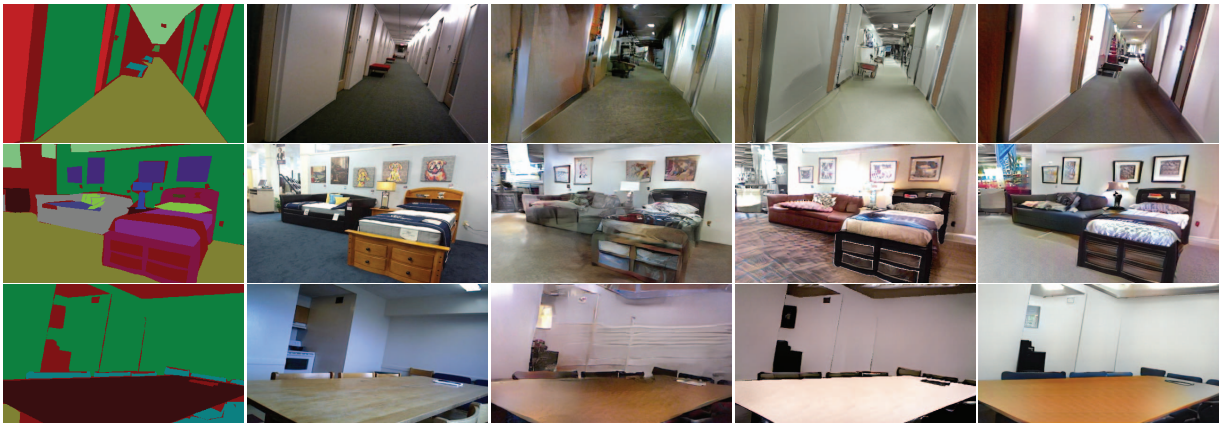
We also qualitatively compare the depth images that are generated by our method from the label maps with the depth images that are estimated by the mono depth estimation method [21] from the real color images in Fig. 5. The results show that our approach can infer reasonable depth maps only from semantic information. A quantitative comparison is given in Table 4.3.

## 4.2. Generating Training Data for RGB-D Semantic Segmentation

In Table 4.2, we validate that an RGB-D semantic segmentation method can benefit from our generated data significantly. This is in particular relevant for robotics applications where annotated training data is limited. In this case, our approach can generate more variants for each annotated training image. Specifically, we train our method on the NYUv2 [26] training set and generate new RGB-D images for each semantic layout in the training set. We then mix the generated and original image by randomly picking some semantic classes and replacing the pixels for these classes in the original image with the pixels of the generated image as shown in Fig. 6. This generates the same dataset. Finally, we train the RGB-D semantic segmentation method ESANet [25] on the mixed images and evaluate the mIoU on the real images of the test set.

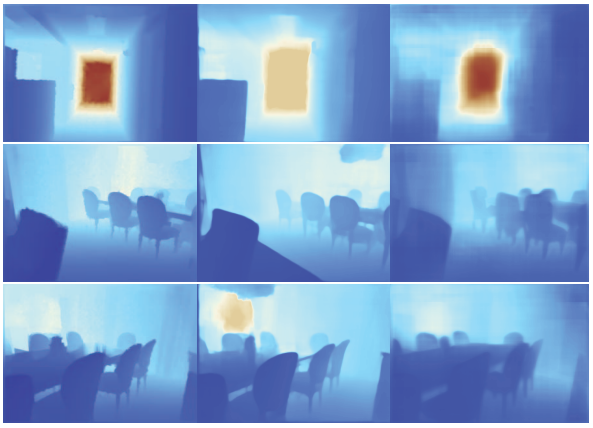
Table 4.2 shows the results for different percentages of randomly replaced classes. We furthermore evaluate if only depth, RGB values, or RGB-D values are replaced. The results show that all settings boost the performance significantly (4.89 - 6.16 mIoU), which demonstrates that the generated data increases the diversity of the dataset and improves the training of the baseline. The best results are achieved by replacing only RGB values, which is expected since appearance has more diversity than depth. Nevertheless, even replacing only depth values significantly improves the segmentation accuracy.

Since the best performance is achieved by only replacing RGB values, we also compare the results when we generate the data with a uni-modal generator. For this experiment, we use the setting where we replace 70% of the classes. The results in Table 4.2 show the benefit of multi-modal



(a) Label (b) Real (c) DAGAN [29] (d) OASIS [24] (e) SCMIS (Ours)

Figure 4. Qualitative results of generated RGB images on the NYU dataset. The two most left columns show the input label map (a) and the corresponding real image (b). Compared to DAGAN and OASIS, our method can generate more realistic images. Besides better details and fewer artefacts, the illumination also looks more realistic due to the estimated geometry.

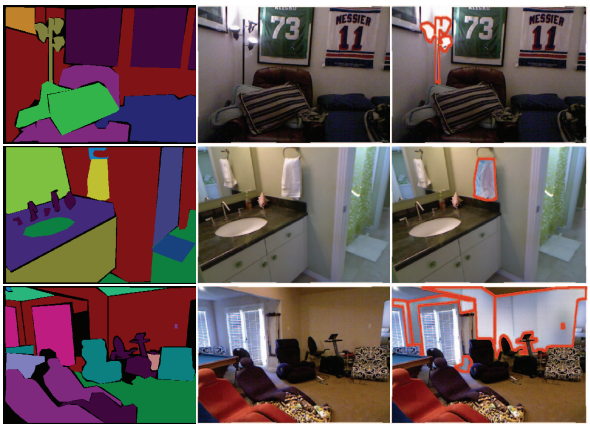


(a) Real (b) Estimation [21] (c) Generation (Ours)

Figure 5. Comparison of the real depth image (a) with the depth map that is estimated by [21] from the real color image (b) and the depth image that is generated by our method from the label map (c).

| Modality | Ratio | mIoU ↑ |
|---|---|---|
| - | 0.0 | 42.08 |
| Depth | 0.3 | 47.40 (+5.32) |
| | 0.5 | 47.57 (+5.49) |
| | 0.7 | 47.77 (+5.69) |
| RGB | 0.3 | 47.77 (+5.69) |
| | 0.5 | 47.71 (+5.63) |
| | 0.7 | **48.24 (+6.16)** |
| RGB-D | 0.3 | 46.97 (+4.89) |
| | 0.5 | 47.85 (+5.77) |
| | 0.7 | 47.84 (+5.76) |

Table 2. Impact of mixing generated and real images on the NYUv2 dataset with image resolution 256x512. '-' denotes the baseline trained without generated images.



(a) Label (b) Real (c) Mixed

Figure 6. Visualization of the semantic segmentation label (a) with the real color image (b) and the real and generated mixed color image (c), we highlight the generated pixels with red contour.

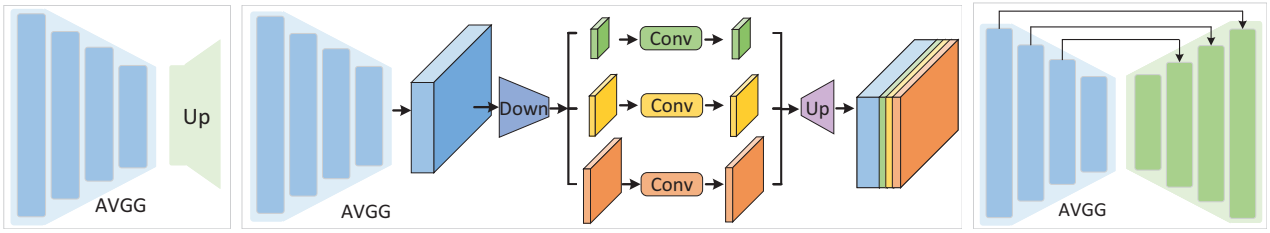
| Methods | OASIS [24] | SCMIS (Only RGB) | SCMIS |
|---|---|---|---|
| mIoU ↑ | 47.56 | 47.67 | **48.24** |

Table 3. Impact of different generation methods on the NYUv2 dataset. SCMIS (Only RGB) denotes SCMIS without geometry decoder.

image generation. When we use the uni-modal method OASIS [24] or train SCMIS in a uni-modal setting, i.e., without the geometry decoder, the segmentation accuracy is lower. This is consistent with Table 4.3, which shows that the multi-modal approach generates more realistic images, i.e., lower FID score.

### 4.3. Ablation Study

In this section, we analyse different design choices of the proposed approach. The experiments include RGB-D




(a) Upsample (b) Pyramid Pooling (c) U-Net

Figure 7. Three architectures that are investigated for the discriminator. Each network uses an adaptive VGG-19 backbone (AVGG).

| Methods | NYU | | Cityscapes | | Size |
|---|---|---|---|---|---|
| | FID ↓ | mIoU$_D$ ↑ | FID ↓ | mIoU ↑ | (MB) |
| OASIS [24] | 77.5 | 58.3 | 47.7 | 69.3 | 22.2 |
| Upsample | 63.7 | 58.0 | 51.1 | 71.9 | **4.1** |
| UNet | 59.6 | **59.6** | 48.2 | **73.1** | 5.1 |
| PP | **55.7** | 59.2 | **43.2** | 72.2 | 6.5 |
| RGB-D PP | 77.3 | 55.9 | N/A | N/A | 6.5 |

Table 4. Impact of different discriminator architectures. PP denotes Pyramid Pooling and RGB-D PP is a discriminator with Pyramid Pooling but takes RGB-D images as input.

| Methods | Cityscapes | |
|---|---|---|
| | FID ↓ | mIoU ↑ |
| PP | 43.2 | **72.2** |
| PP + $\mathcal{L}_{A_p}$ (IN) | 43.4 | 72.1 |
| PP + $\mathcal{L}_{A_p}$ (CS) | 42.1 | 72.1 |
| PP + $\mathcal{L}_{A_{ap}}$ | **41.7** | 72.1 |

Table 5. Impact of perceptual loss. $\mathcal{L}_{A_p}$: perceptual loss with features pre-trained on ImageNet (IN) or Cityscapes (CS); $\mathcal{L}_{A_{ap}}$: adaptive perceptual loss.

| Methods | Modality | NYU | | | | |
|---|---|---|---|---|---|---|
| | | FID ↓ | mIoU$_D$ ↑ | AbsRel ↓ | RMSE ↓ | SqRel ↓ |
| SCMIS | Depth | N/A | N/A | 0.172 | 0.613 | 0.142 |
| SCMIS | RGB | 59.2 | N/A | N/A | N/A | N/A |
| RGB-D OASIS | RGB-D | 89.1 | 54.7 | 0.221 | 0.737 | 0.204 |
| SCMIS | RGB-D | 55.7 | 59.2 | **0.166** | **0.599** | **0.133** |
| 4D Gen + $\mathcal{L}_{A_{ap}}$ | RGB-D | 65.8 | **60.0** | 0.197 | 0.662 | 0.170 |
| SCMIS + $\mathcal{L}_{A_{ap}}$ | RGB-D | **55.2** | 59.8 | 0.167 | **0.599** | 0.134 |

Table 6. Comparison of uni-modal and multi-modal generators.

image synthesis (NYU dataset) and RGB image synthesis (Cityscapes dataset). For RGB image synthesis, we use the uni-modal generator OASIS [24] in combination with our discriminator.

**Architectures of Discriminator.** In Table 4.3, we show the impact of different architectures on the discriminator. The architectures Upsample, U-Net, and Pyramid Pooling (PP) are shown in Fig. 7. For this experiment, we do not use adaptive perceptual loss. Without multi-scale ability (Upsample), the quality of the generated images is relatively low (high FID). Using multi-scale architectures for the discriminator clearly improves the results (U-Net and PP). Compared to U-Net, Pyramid Pooling improves the quality of the generated images by nearly 4-5 FID at the cost of a slightly lower semantic consistency (less than 1 mIoU). We thus use Pyramid Pooling (PP) for our discriminator. So far, the discriminator has only taken the RGB image as input, but not the depth map. We, therefore, evaluate what happens if we provide the RGB-D images as input to the discriminator. We denote this setting by RGBD-PP. We can see that both image quality and semantic consistency drop drastically. This shows that considering depth and color as the same modality for the discriminator does not work well. Compared to the discriminator of OASIS, our discriminator is much more compact (nearly 4× smaller) and it results in more realistic images for both RGB-D image synthesis (FID is reduced more than 20) and RGB image synthesis (FID is reduced more than 4).

**Adaptive Perceptual Loss.** So far, we have not used the adaptive perceptual loss (4). We evaluate its impact for semantic RGB image synthesis in Table 4.3. The adaptive perceptual loss ($\mathcal{L}_{A_{ap}}$) improves the image quality and reduces FID by 1.5, while mIoU remains the same. This is expected since the loss does not measure the semantic consistency with the label map, but the perceptual similarity to the real image. If we use the standard perceptual loss ($\mathcal{L}_{A_p}$ (IN)), the FID score even slightly increases. This is consistent with our observation that it pushes the generated features towards a feature space where the classes are not separable and difficult to classify as it is illustrated in Fig. 3. This can be also addressed by training the VGG-19 features on the training set before training the generator ($\mathcal{L}_{A_p}$ (CS)). While this reduces the FID score, it does not perform as well as the adaptive perceptual loss. Furthermore, it requires training two separate networks, which is not very practical.

**Multi-Modality.** In Table 4.3, we compare our multi-modal approach (row 4) to the uni-modal variants where we estimate the depth maps (row 1) and the RGB images (row 2) by separately trained networks. We use Absolute Relative Error (AbsRel), Root Mean Square Error (RMSE), and Square Relative Error (SqRel) [3] to measure the accuracy of the depth maps. Both the image quality and the accuracy of the depth maps are higher if our network is trained for both modalities. This shows that depth improves the image

quality and that appearance improves the generated depth maps.

**Architectures of Generator.** Furthermore, we compare our approach to an extension of the uni-modal method OASIS [24] to RGB-D images (RGB-D OASIS). We use the generator from OASIS but modify it such that it generates RGB-D images. For a fair comparison, we use the discriminator and loss functions from our approach since an RGB-D discriminator does not perform well as we showed in Table 4.3. Note that we do not use the adaptive perceptual loss for these experiments. Compared to our method, RGBD-OASIS (row 3 in Table 4.3) performs much worse both in terms of image quality and depth accuracy. A few qualitative results are shown in Fig. 1.

Finally, we compare our generator with two modal-dependent decoders for appearance and geometry to a variant with a single decoder for appearance and geometry (4D Gen) in Table 4.3 (row 5). While the semantic consistency (mIoU) is the same, the quality of the generated RGB and depth images is lower compared to our approach.

## 5. Conclusion

In this paper, we proposed an approach to increase the appearance diversity of an annotated dataset for RGB-D semantic segmentation. To this end, we addressed the new task of semantic RGB-D image synthesis and proposed a semantic consistent multi-modal generator, which comprises a modal-independent encoder and two modal-dependent decoders. It fully utilizes multi-modal information and produces semantic consistent and realistic RGB-D images. We have demonstrated that the proposed approach performs by a large margin better than uni-modal approaches and that it can be used to improve the accuracy of an RGB-D semantic segmentation method. While the experimental evaluation focused on semantic segmentation, we expect that the approach can also be used for other robotics-related tasks where the number of training images is limited.

## References


[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan, 2017. 2

[2] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 5

[3] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *NeurIPS*, 2014. 7

[4] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, Kun Zhang, and Dacheng Tao. Geometry-consistent generative adversarial networks for one-sided unsupervised domain mapping. In *CVPR*, 2019. 2

[5] Baris Gecer, Stylianos Ploumpis, Irene K, and Stefanos Zafeiriou. Ganfit: Generative adversarial network fitting for high fidelity 3d face reconstruction. In *CVPR*, 2019. 2

[6] Baris Gecer, Stylianos Ploumpis, Irene Kotsia, and Stefanos Zafeiriou. Fast-ganfit: Generative adversarial network for high fidelity 3d face reconstruction. In *TPAMI*, 2021. 2

[7] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. In *arXiv*, 2014. 2

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3

[9] 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 *arXiv*, 2017. 5

[10] Tao Hu, Liwei Wang, Xiaogang Xu, Shu Liu, and Jiaya Jia. Self-supervised 3d mesh reconstruction from single images. In *CVPR*, 2021. 2

[11] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. 2

[12] Liming Jiang, Changxu Zhang, Mingyang Huang, Chunxiao Liu, Jianping Shi, and Chen Change Loy. Tsit: A simple and versatile framework for image-to-image translation. In *ECCV*, 2020. 2, 5

[13] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016. 2, 4

[14] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, 2012. 3

[15] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *ICCV*, 2017. 2

[16] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. In *arXiv*, 2014. 2

[17] Xingang Pan, Bo Dai, Ziwei Liu, Chen Change Loy, and Ping Luo. Do 2d gans know 3d shape? unsupervised 3d shape reconstruction from 2d image gans. In *arXiv*, 2020. 2

[18] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *CVPR*, 2019. 1, 2, 4, 5

[19] Jingtan Piao, Keqiang Sun, Quan Wang, Kwan-Yee Lin, and Hongsheng Li. Inverting generative adversarial renderer for face reconstruction. In *CVPR*, 2021. 2

[20] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *arXiv*, 2015. 2

[21] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *ICCV*, 2021. 5, 6

[22] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 4

[23] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *arXiv*, 2016. 2

[24] Edgar Schönfeld, Vadim Sushko, Dan Zhang, Juergen Gall, Bernt Schiele, and Anna Khoreva. You only need adversarial supervision for semantic image synthesis. In *ICLR*, 2021. 1, 2, 4, 5, 6, 7, 8

[25] Daniel Seichter, Mona Köhler, Benjamin Lewandowski, Tim Wengefeld, and Horst-Michael Gross. Efficient rgb-d semantic segmentation for indoor scene analysis. In *ICRA*, 2021. 1, 5

[26] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 5

[27] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *CVPR*, 2015. 5

[28] Vadim Sushko, Edgar Schönfeld, Dan Zhang, Juergen Gall, Bernt Schiele, and Anna Khoreva. OASIS: only adversarial supervision for semantic image synthesis. In *IJCV*, 2022. 1

[29] Hao Tang, Song Bai, and Nicu Sebe. Dual attention gans for semantic image synthesis. In *ACM MM*, 2020. 2, 5, 6

[30] Hao Tang, Dan Xu, Yan Yan, Philip HS Torr, and Nicu Sebe. Local class-specific and global image-level generative adversarial networks for semantic-guided scene generation. In *CVPR*, 2020. 2

[31] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. In *arXiv*, 2016. 2

[32] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. In *JMLR*, 2008. 3

[33] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *CVPR*, 2018. 2

[34] Wayne Wu, Kaidi Cao, Cheng Li, Chen Qian, and Chen Change Loy. Transgaga: Geometry-aware unsupervised image-to-image translation. In *CVPR*, 2019. 2

[35] Sheng Xu, Yanjing Li, Teli Ma, Mingbao Lin, Hao Dong, Baochang Zhang, Peng Gao, and Jinhu Lu. Resilient binary neural network. In *AAAI*, 2023. 2

[36] Sheng Xu, Yanjing Li, Bohan Zeng, Teli Ma, Baochang Zhang, Xianbin Cao, Peng Gao, and Jinhu Lü. Ida-det: An information discrepancy-aware distillation for 1-bit detectors. In *ECCV*, 2022. 2

[37] Fisher Yu, Vladlen Koltun, and Thomas Funkhouser. Dilated residual networks. In *CVPR*, 2017. 5

[38] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017. 4

[39] Xinge Zhu, Zhichao Yin, Jianping Shi, Hongsheng Li, and Dahua Lin. Generative adversarial frontal view to bird view synthesis. In *3DV*, 2018. 2